Universidade Federal de Santa Catarina

Programa de Pós-Graduação em Engenharia de produção

APLICAÇÃO DE DATA WEBHOUSING PARA MONITORAMENTO DE ACESSOS A SITES WEB DE GRUPOS DE PESQUISA E DESENVOLVIMENTO: Um Estudo de Caso

Daniel Martins Barbosa

Dissertação apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do grau de Mestre, em
Engenharia de Produção

Prof. Dr. Roberto C. S. Pacheco
Orientador

Florianópolis
2003

DANIEL MARTINS BARBOSA

APLICAÇÃO DE DATA WEBHOUSING PARA MONITORAMENTO DE ACESSOS A SITES WEB DE GRUPOS DE PESQUISA E DESENVOLVIMENTO: Um Estudo de Caso

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA.
ESPECIALIDADE EM ENGENHARIA DE PRODUÇÃO E APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

_____
Edson Pacheco Paladini, Dr
Coordenador do Programa de Pós-Graduação em Engenharia de Produção

BANCA EXAMINADORA

_____
Roberto C. S. Pacheco, Dr. - Orientador

_____
Denílson Sell, Doutorando, Msg. - Tutor

_____
José Leomar Todesco, Dr.

_____
Tânia Fátima Calvi Tait, Dra.

AGRADECIMENTOS

A DEUS, NOSSO SENHOR JESUS CRISTO e ao ESPIRITO SANTO, que me iluminaram e me deram saúde, sabedoria e força de vontade para não desanimar no meio do caminho. Por terem-me dado a oportunidade de realizar um sonho há tanto tempo acalentado.

À minha família, de maneira muito especial e carinhosa às três pessoas pelas quais tenho grande amor e carinho, que souberam esperar, pacientemente, durante mais esta jornada, compreendendo as minhas ausências e proporcionando apoio diante dos maus momentos sem, em nenhum instante, deixar de incentivar-me: Elizete, minha querida esposa, e nossos dois filhos Bruno e Fernanda, alegria de nossas vidas. Sem vocês seria impossível chegar até aqui.

Ao meu orientador, professor Roberto Carlos dos Santos Pacheco, por ter aceitado minha proposta de dissertação e por ter contribuído de modo eficiente e eficaz para consolidação deste trabalho.

Agradeço também aos amigos e tutores que de muitas maneiras participaram deste trabalho, dando sugestões valiosas. Sou muito grato em especial a Denílson Sell, Tânia Fátima Calvi Tait e Olival de Gusmão Freitas Junior.

À Universidade Estadual de Maringá, especialmente aos amigos do Núcleo de Processamento de Dados (NPD), Departamento de Informática (DIN) e Pró-Reitoria de Pesquisa e Pós-Graduação (PPG), pela amizade e pelo apoio em toda esta caminhada.

Aos Professores e amigos, Álvaro José Periotto, Dante Alves de Medeiro Filho e Moacir José da Silva que confiaram em mim, apresentando-me ao Programa de Pós-Graduação e pela valiosa colaboração, abrindo o caminho para a reta final deste trabalho.

Agradeço, igualmente, aos meus grandes amigos Altevir, Edinaldo, Enee Saulo e suas famílias, sempre presentes quando precisei.

E, gostaria de agradecer também aos meus pais, João M. Barbosa Filho e Ostrilia da S. Barbosa, pelas orações, incentivo, confiança e exemplos de força e dedicação. Pessoas nas quais me espelho, constantemente.

Finalmente, aos meus amigos de toda a vida, em especial aos amigos do Grupo Stela que, com sua presença, pequenos gestos ou palavras, contribuíram para a minha chegada a mais esta etapa.

“Há homens que lutam um dia e são bons,

há homens que lutam muitos dias e são melhores,

porém, há os que lutam toda a vida esses são os imprescindíveis”

(Bertold Brecht).

Basta-nos tentar, tentar...

RESUMO

BARBOSA, Daniel Martins, **Aplicação de Data Webhousing para Monitoramento de Acessos a Sites Web de Grupos de Pesquisa e Desenvolvimento: Um Estudo de Caso**. 123f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção. UFSC, Florianópolis - 2003.

O presente Trabalho tem como objetivo utilizar as técnicas de *Data Warehouse* para extrair informações, bem como propor um *Data Mart* para monitorar *sites* de Grupos de Pesquisa e Desenvolvimento. Nesse sentido, ao marcarem presença na *Web*, as organizações e os grupos de pesquisa estão abrindo um canal extremamente poderoso para a criação e o desenvolvimento de relações com seus usuários. Assim, é necessário que essas organizações incorporem ferramentas para monitorar seus usuários e suas ações, visando atender à necessidade de conhecer o comportamento dos seus usuários. Como resultado desse novo contexto surge o compromisso de criação de uma nova metodologia que priorize o tratamento de informações, tanto estruturadas, quanto não estruturadas.

Este trabalho apresenta o desenvolvimento e a aplicação das técnicas de *Data Webhouse* que visa fornecer subsídios à coleta e análise de dados para *site* de grupo de P&D. Levando-se em consideração esse cenário, será apresentada uma metodologia que permite as especificações de variáveis de um processo de personalização, por ser uma aplicação que auxilia na estruturação de *sites* de grupos de P&D.

Dessa forma, conclui-se que a utilização dessa metodologia possibilitará uma reestruturação nos *sites* e apresentar uma contribuição em que se destaca um modelo geral de *site* para Grupos de P&D.

Palavras-chave: DW, *Data Webhousing*, Grupos de P&D, Sistemas de Apoio à Decisão e *WebSite*.

ABSTRACT

BARBOSA, Daniel Martins, **Aplicação de Data Webhousing para Monitoramento de Acessos a Sites Web de Grupos de Pesquisa e Desenvolvimento: Um Estudo de Caso**. 123f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção. UFSC, Florianópolis, 2003.

This study aims at using Data Warehouse techniques to extract information, as well as proposing a Data Mart to monitor sites of Development and Research Groups. In this sense, being in the Web, the organizations and the research groups are opening an extremely powerful channel for the creation and the development of relationships with its users. Therefore, it is necessary that such organizations incorporate tools to monitor their users and their own actions, with the purpose of knowing their users' behavior. As a result of that new context, the commitment to create a new methodology that prioritizes the treatment of both structured and non structured information appears.

This study the Data Webhouse's techniques development and application aiming at supplying with subsides to data collection and analysis for the P&D Group site. Considering this scenery, a methodology that allows the specifications of a personalization process variables is presented, since it is an application that helps with the P&D sites structuring.

Therefore, we can conclude that the use of such a methodology will make the sites restructuring possible, as well as it will contribute with a general site model for P&D Groups.

Key-words: DW, Data Webhousing, Groups of P&D, Systems of Support the Decision and WebSite**.**

SUMÁRIO

Lista de Figuras

Lista de Quadros

Siglas

ASP	**ACTIVE SERVER PAGES**
BD	**BANCO DE DADOS**
BI	**BUSINESS INTELLIGENCE**
C&T	**CIÊNCIA E TECNOLOGIA**
CFL	**COMMON LOG FORMAT**
CGI	**COMMON GATEWAY INTERFACE**
CNPQ	**CONSELHO NACIONAL DE PESQUISA**
CRM	**CUSTOMER RELATIONSHIP MANAGEMENT**
CT&I	**CIÊNCIA E TECNOLOGIA E INFORMAÇÃO**
DHTML	**LINGUAGEM DE MARCAÇÂO DE HIPERTEXTO DINAMIC**
DM	**DATA MART**
DM	**DATA MINING**
DW	**DATA WAREHOUSE**
DW	**DATA WEBHOUSE**
ETL	**EXTRAÇÃO TRANSFORMAÇÃO E CARGA**
HTML	**LINGUAGEM DE MARCAÇÂO DE HIPERTEXTO**
HTTP	**PROTOCOLO DE TRANSFERENCIA DE HIPERTEXTO**
IA	**INTELIGENCIA ARTIFICIAL**
ID	**IDENTIFICAÇÂO DE SESSÃO**
IP	**PROTOCOLO DE INTERNET**
ISBN	**INTERNATIONAL STANDARD BOOK NUMBER**
JSP	**JAVA SERVER PAGES**
MOLAP	**MULTIDIMENSIONAL ON-LINE ANALYTICAL PROCESSING**
KDD	**KNOWLEDGE DISCOVERY IN DATABASE**
ODS	**SERVIDOR DE ARMAZENAMENTO DE DADOS OPERACIONAIS**
OLAP	**ON LINE ANALYTICAL PROCESSING**
OLTP	**ONLINE TRANSACTION PROCESSING**
P&D	**PESQUISA E DESENVOLVIMENTO**
PPGEP	**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**
RBC	**RACIOCÍNIO BASEADO EM CASO**
ROI	**RETORNO DE INVESTIMENTO**
ROLAP	**RELATIONAL ON-LINE ANALITICA PROCESSING**
SAD	**SISTEMA DE APOIO À DECISÃO**
SGBD	**SISTEMA GERAL DE BANCO DE DADOS**
SSL	**SECURE SOCKETS LAYER**
TI	**TECNOLOGIA DE INFORMAÇÃO**
UFSC	**UNIVERSIDADE FEDERAL DE SANTA CATARINA**
UEM	**UNIVERSIDADE ESTADUAL DE MARINGÁ**
URL	**LOCALIZADOR UNIFORME DE RECURSOS.**
XML	**EXTENSIBLE MARKUP LANGUAGEM**

1 INTRODUÇÃO

1.1 Contexto e Relevância do Problema

Já é notório o fato de que a rápida evolução nas tecnologias e as mudanças organizacionais tornam fundamental aos dirigentes das organizações possuírem grande rapidez nas suas decisões. Assim Sendo, eles necessitam de informações atuais, confiáveis e precisas, que devem ser correlacionadas de tal forma que lhes permitam tomar decisões mais facilmente e trabalhar com cenários futuros (DALFOVO e FRANCO, 2000).

Atualmente, na área de informática, surgem novas tecnologias que integram num único sistema todas as informações necessárias para que o executivo possa verificá-las de forma rápida e confiável, desde o nível operacional até o nível mais analítico que se desejar, possibilitando-lhe um maior conhecimento e controle da situação, além de maior agilidade e segurança no processo decisório (FURLAN, 1994).

Para Dalfovo e Franco (2000), a demanda por informações estratégicas de apoio à decisão tem sido um dos grandes motivos para investimentos em soluções informatizadas. Com os avanços tecnológicos, as organizações atingiram um nível satisfatório de informatização sobre seus processos operacionais. O próximo passo é transformação do grande volume de dados gerado pelos diversos sistemas de informação útil e acessível aos tomadores de decisão.

Segundo Dias (2001), é imprescindível a aplicação da gestão da informação para administrar o caos informacional do mundo digital. Muitas vezes, essas informações encontram-se armazenadas em bases de dados não integradas e em plataformas de sistemas operacionais e gerenciadores de banco de dados diferentes. Desse modo, o acesso a tais informações torna-se difícil e, conseqüentemente, o processo de tomada de decisão também é dificultado, considerando-se uma visão global da organização.

Para Sell (2001), atualmente, a tecnologia da maioria dos novos projetos de sistemas de informação foi ou está sendo desenvolvida por meio de uma técnica denominada *Data Warehousing*, cujo propósito é a concepção de sistemas baseados na estruturação de um repositório de dados informacionais.

De acordo com Inmon (1997), um sistema de *Data Warehouse* é composto, além de outras ferramentas, de um banco de dados para o qual somente as informações necessárias

para a tomada de decisão são carregadas, as quais são oriundas de sistemas de informações operacionais. Como esse novo banco de dados contém apenas as informações essenciais, as pesquisas feitas nele são rápidas e podem responder questões a complexas.

Segundo Cielo (2001), entre as tecnologias existentes no mercado, o *Data Warehouse* constitui-se ferramenta adequada para solucionar as dificuldades de integração de dados mantidos em plataformas tecnológicas divergentes. Ferreira (2001), destaca que, devido à grande expansão da Internet e à proliferação de portais corporativos, surgiu uma nova técnica denominada *Data Webhouse,* derivada do *Data Warehousing* e dos conceitos de publicação de informações introduzidas pelos *sites* na *web*.

Para Kimball (2000), *Data Webhouse* é a instanciação da *web* para o *Data Warehouse*. Para esse autor, trazer a *web* para o *Warehouse* significa trazer comportamentos de uso das informações para o *Data Warehouse*. Dessa forma, as seqüências de cliques (*clickstream*) representam um registro potencial do comportamento dos usuários na *web* e o objetivo do *Data Webhouse* é capturar, analisar e entender esse comportamento.

Com essa tecnologia, é possível acompanhar o relacionamento dos usuários dos *sites* dos Grupos de Pesquisa, o que pode subsidiar a personalização destes *sites* de acordo com o perfil de utilização do usuário. Essas informações também podem oferecer procedimentos de reorganização da estruturação dos *sites* e das informações veiculadas, além de possibilitarem a redefinição de uma política de integralização dos grupos de pesquisas.

Considerando-se esse novo cenário, observa-se que a utilização dessas tecnologias pode facilitar a tomada de decisão do gestor no processo decisório, visto que possibilita a obtenção de dados e de conteúdos relevantes para as organizações das mais diversas áreas. Por essa razão, o presente trabalho tem por objetivo aplicar técnicas de *Data Warehousing* e, principalmente, *Data Webhousing* ao *site* de um grupo de pesquisa e desenvolvimento, igualmente apresentar um *Data Mart Clickstream*[1] como um estudo de caso que surge da integração dessas tecnologias.

A questão de pesquisa deste trabalho refere-se à possibilidade de que a técnica de *Data Webhousing* possa contribuir com projetos de *sites web* de grupos de pesquisa e desenvolvimento (P&D). *Sites Web* constituem uma das principais formas pelas quais Grupos de Pesquisa e desenvolvimento publica resultados de pesquisa e apresentam seu portfólio de

[1] Pequenos conjuntos de Data Webhouse baseados em assuntos específicos (Inmon, 1997 e Kimball, 2000).

soluções, além de, nos casos de Grupos de Pesquisa ligados à área tecnológica, servirem de meio de publicação de serviços de informação. A melhoria nesses objetivos pode aumentar a visibilidade dos Grupos de P&D, potencialmente ampliando suas atividades de intercâmbio.

Os resultados discutidos ao final da dissertação incluem indicadores que podem contribuir para que os Grupos de Pesquisa conheçam melhor os usuários de seu *site* institucional, permitindo assim que novas possibilidades de serviços sejam viabilizadas.

Dentro dessa perspectiva, a pergunta de pesquisa que surge e que motiva o presente estudo é a seguinte: *quais são os indicadores para análise dos sites de Grupos de P&D que possibilitará contribuições em sites num ambiente de Ciência e Tecnologia (C&T)*?

1.2 Objetivo Geral

Aplicar a teoria de *Data Webhouse* no desenvolvimento de um ambiente de análise dos acessos ao *site* de um Grupo de Pesquisa e Desenvolvimento Científico, Acadêmico e Tecnológico.

1.2.1 Objetivos Específicos

- Fazer um levantamento sobre *Data Webhousing*, envolvendo os principais conceitos e características relacionados à técnica de como gerenciar *Data Webhouse.*
- Apresentar um levantamento bibliográfico sobre as metodologias de *Data Webhousing.*
- Implementar um protótipo de *Data Mart* de seqüência de clique, utilizando *Data Webhousing.*
- Propor melhorias aos *websites* de Grupos de Pesquisa e Desenvolvimento, através da análise dos resultados obtidos sobre o protótipo desenvolvido para um *site* e propor uma rotina de reestruturação, bem como também apresentar um modelo *sites* para Grupos de P&D.

1.3 Justificativa do Trabalho

A Internet apresenta-se cada vez mais como uma nova mídia para comunicação entre as empresas e o mercado. No entanto, ainda não está claro em que grau ela tem contribuído para a alavancagem de negócios nas organizações que a utilizam.

Segundo Kimball (2000f), há uma constatação geral de que a Internet "é a próxima fronteira" da competividade, entretanto, as iniciativas ainda são poucas para a maioria das organizações, que estão sendo levadas mais pelo desejo de acompanhar a inovação tecnológica do que por explorar novas alternativas de negócios.

Nesse contexto, as ações do usuário da *web* podem ser motivadas por um número de necessidades diferentes que podem mudar de um momento para outro durante uma sessão de visita a um *web site.* Assim, as organizações podem usar essas informações para determinar hábitos de compra, fornecer aos usuários recomendações sobre novos produtos, entre outros. Todavia, é necessária a utilização de uma infra-estrutura tecnológica para uma avaliação eficaz desses dados.

Dessa forma, Kimball (2000f) ressalta que o *Data Webhouse* pode tornar-se o elemento central e coesivo das instituições, fornecendo informações competitivas e essenciais aos pesquisadores e responsáveis pelas decisões estratégicas. Por isso, encontrar o perfil das pessoas que acessam *sites na web*, tornou-se necessário, devido à grande importância dada pelos pesquisadores ao conhecimento de seus usuários, e o *Data Webhouse* possui as técnicas adequadas para o gerenciamento dos relacionamentos entre eles.

Para Barbosa *et al*., (2002), o objetivo do *Data Webhouse* é capturar, armazenar e publicar dados de seqüência de clique que possibilitam uma compreensão do comportamento do usuário na *web.* Se os administradores possuem informações sobre cada clique de seus usuários, ou seja, o caminho percorrido por eles dentro do *web site,* certamente por meio das técnicas e ferramentas de *Data Webhouse*, podem ter informações muito relevantes para a tomada de suas decisões. Dessa forma então para se obter êxito com estas informações o *site* precisa ser bem projetado, devendo-se incorporar ferramentas para monitorar as ações dos visitantes, ou seja, é preciso que ele funcione como uma forma de medir a eficácia[2] do *site.*

[2] Eficácia. consiste da capacidade de pessoas e instituições alcançarem objetivos e metas, ou seja, os resultados com os quais se comprometeram ou a que foram propostos. Em outras palavras o fazer certo. (Campos, 2000).

Assim, tais informações de monitoramento podem ser utilizadas para personalizar o conteúdo dinâmico do *site,* aumentando a relevância desse conteúdo à medida que a sessão progride ou quando o usuário retorna ao *site.* Dessa maneira, as pessoas conhecerão melhor quem acessa seu *site*, ou seja, quem são os seus usuários.

Grupos de Pesquisa e Desenvolvimento utilizam a Internet para divulgação de seus trabalhos acadêmicos e científicos, para divulgação de suas áreas de pesquisa, para promoção de intercâmbio e cooperação com outros Grupos Científicos e para apresentação de sua imagem institucional. Com as técnicas de *Data Webhousing*, pode-se verificar o grau com que o *web site* de um Grupo de Pesquisa está alcançando seus objetivos. Como resultado, abrem-se perspectivas de melhoria do *web site* dos Grupos de Pesquisa e, conseqüentemente, uma melhor configuração do conteúdo e das possibilidades de serviços desses Grupos, através de um novo *site.*

1.4 Metodologia

Apresenta-se a seguir, a classificação deste trabalho quanto à sua natureza, sua forma de abordar o problema e seus objetivos, bem como os procedimentos técnicos utilizados. Quanto ao primeiro parâmetro, isto é, a natureza, o presente trabalho classifica-se como "pesquisa aplicada", pois "objetiva gerar conhecimentos para aplicação prática dirigindo soluções de problemas específicos" (Silva e Menezes, 2001, p. 20), envolvendo verdades e interesses locais. Do ponto de vista da forma de abordagem do problema, a pesquisa caracteriza-se como predominantemente qualitativa, uma vez que requer o uso de recursos e de técnicas estatísticas (Silva e Menezes, 2001. p. 20).

Segundo Gil (1991), quanto aos objetivos, a pesquisa procura identificar o problema com vistas a torná-lo explícito ou a construir hipóteses. Para alcançá-lo, faz-se levantamento bibliográfico, entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado e analisam-se exemplos que estimulem a compreensão, em geral procede-se a Pesquisas Bibliográficas e estudos de caso. Essas características qualificam este trabalho como sendo de natureza "exploratória"; devido a procedimentos técnicos utilizados, ele tem características de estudo de caso, em decorrência da determinação dos objetivos.

Do ponto de vista metodológico, optou-se pelo estudo de caso, em função deste método possibilitar uma profundidade e riqueza maior, dando assim mais embasamento à pesquisa.

Esta pesquisa conduziu-se em seis etapas: Revisão Bibliográfica, Sistemas de Apoio a Decisão, Estudo dos Conceitos, Característica e Metodologias Existentes na Literatura de *Data Webhouse;* Estudo de *Sites* dos Grupos de P&D; Construção de um *Data Mart* para Monitorar *Sites* de Grupos de P&D, Propor Melhorias a *Sites* de P&D e a título de sugestão propor um modelo de *site* para os Grupos de P&D.

Para o entendimento adequado da forma de desenvolvimento da pesquisa, são explicitados cada uma das etapas e seus respectivos desdobramentos.

A Figura 1 retrata o pensamento metodológico quanto ao desenvolvimento dessa pesquisa.

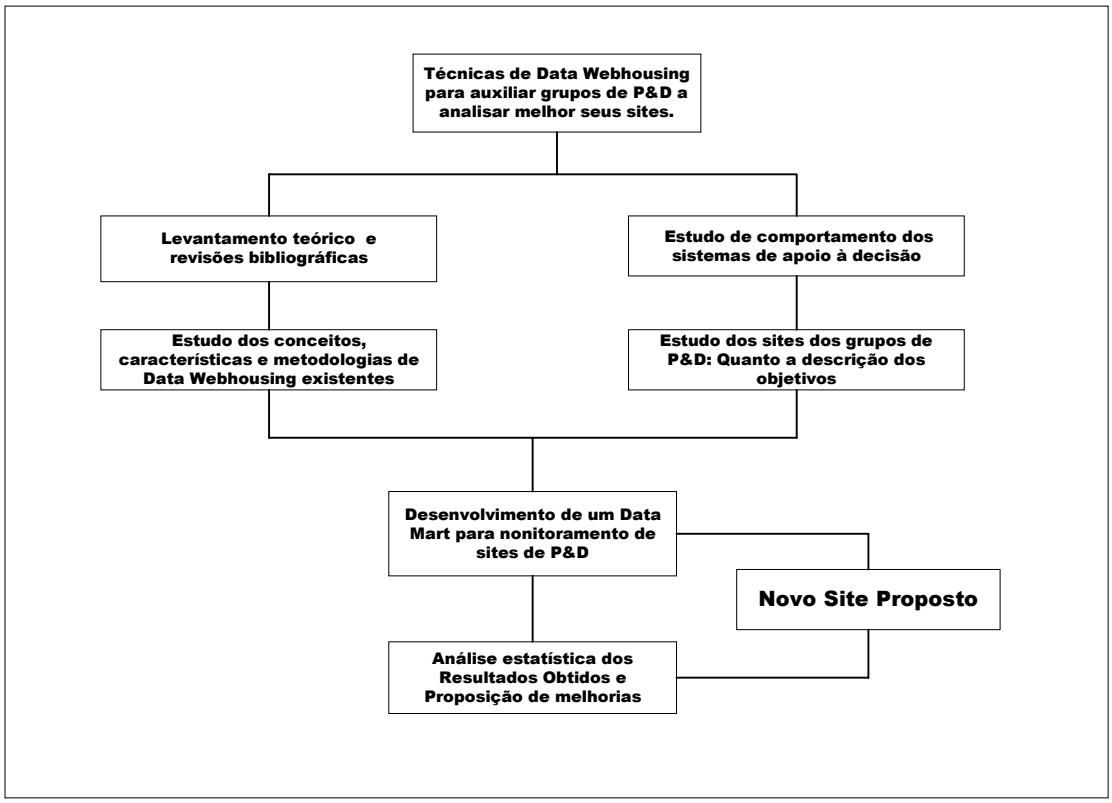


FIGURA 1 – METODOLOGIA UTILIZADA NA PESQUISA

Para atingir os objetivos propostos, desenvolveu-se, inicialmente, um estudo teórico das características, dos problemas e das necessidades de grupos de pesquisa e desenvolvimento, procurando-se evidenciar a importância da informação no processo decisório. E, como um caso de estudo, propor uma aplicação de *Data Webhousing* sobre o *site* de um grupo de pesquisa da Universidade Federal de Santa Catarina e por meio das análises dos indicativos, propor melhorias para os *sites*. Seguem-se as etapas da realização desta pesquisa:

- Etapa 1 – Levantamento teórico sobre as técnicas de *Data Webhousing* - Nesta etapa procurou-se analisar a maioria dos *sites* relacionados à atividade de pesquisa e desenvolvimento, estudando os pontos positivos e negativos de cada página dentro dos *sites*. Observou-se, também, um *site* específico, vale dizer, o *site* do Grupo de Pesquisa Stela da UFSC, bem como se buscou compreender a melhor maneira de relacionar essas estruturas com as técnicas citadas.

- Etapa 2 – Estudo dos Comportamentos dos Sistemas de Apoio à Decisão - Nesta etapa, procurou-se visualizar como ocorre a concepção inicial dos sistemas de apoio à decisão destacando *o Data Warehouse, Data Mining* e o Customer Relationship Management (*CRM*).

- Etapa 3 – Estudo de Metodologias para *Data Webhousing* - O objetivo desta fase foi revisar os conceitos e características da tecnologia Data Webhouse e os principais aspectos de metodologias, principalmente, a proposta de Kimball (2000f), a qual apresenta as fases previstas e as dificuldades de uma implementação de *Data Webhouse.* Foram resgatados, ainda, os processos de desenvolvimento para uma solução voltada ao monitoramento de *web sites*.

- Etapa 4 – Estudo dos *Website* de Grupos de P&D - Nesta fase, buscou-se entender como os *sites web* de pesquisa e desenvolvimento se estruturam, através de um levantamento com alguns *sites* aleatórios.

- Etapa 5 – Desenvolvimento de um *Data Mart* - Esta fase procurou implementar um modelo de *Data Mart* específico como estudo de caso para monitorar o *site* do Grupo de Pesquisa & Desenvolvimento Stela, bem como mostrar os resultados e análise.

- Etapa 6 – Análise de Resultado - Por meio de análise dos indicativos obtidos, propor melhorias para os *sites* dos grupos de P&D, no Brasil.

- Estapa 7 – Proposta de um novo *site* – Através das análises dos indicativos estatísticos e da proposta de melhorias será apresentado um novo *site* para grupos de P&D como modelo referencial.

1.5 Estrutura do Trabalho

Considerando o problema de pesquisa apresentado, bem como os objetivos relacionados a que se refere este trabalho, foi organizado em seis capítulos, além das referências e anexos. A figura 2 apresenta uma visão geral do trabalho.

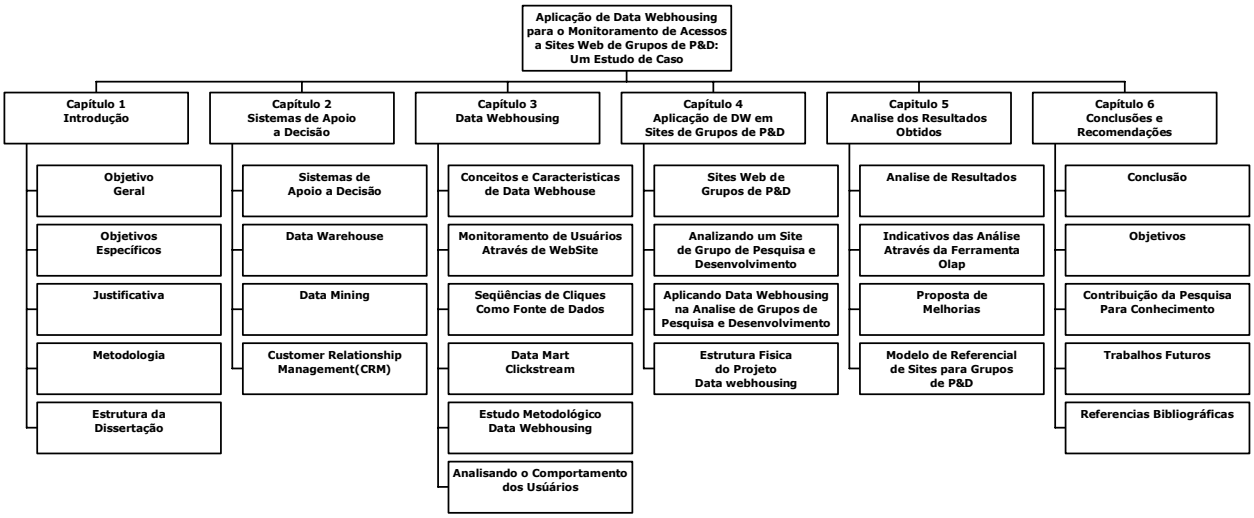


FIGURA 2 – UMA VISÃO GERAL DA ESTRUTURA DA DISSERTAÇÃO.

O primeiro capítulo apresenta os aspectos introdutórios que caracterizam o trabalho, tais como a definição do tema e o problema de pesquisa, o objetivo geral e específico, a sua contribuição teórico-empírica e a descrição da estrutura do trabalho.

O segundo capítulo compreende os conceitos básicos dos Sistemas de Apoio à Decisão, a evolução das ferramentas e as características de cada tecnologia envolvida.

O terceiro capítulo descreve os conceitos e as características do *Data Webhouse*, destacando os estudos das metodologias *Data Webhouse* e o monitoramento de *sites*, procurando mostrar as dificuldades e a complexidade que se tem ao implantar uma solução para análise dos *logs* de *web sites*.

O quarto capítulo expõe o desenvolvimento de um *Data Mart* para monitorar *sites* de Grupos de Pesquisa e Desenvolvimento, procurando relacionar os *sites* de P&D com o uso dessa Tecnologia.

O quinto capítulo apresenta as análises dos resultados obtidos com a implementação descrita no quarto capítulo. E essas análises permitirão sugerir as melhorias, como também

propor um modelo de *site* que venha ao encontro dos objetivos de um *site* de Grupo de Pesquisa e Desenvolvimento.

No sexto capítulo, expõem-se as conclusões do trabalho em resposta aos questionamentos da pesquisa e evidenciam-se recomendações a estudos futuros sobre o tema pesquisado.

2 SISTEMAS DE APOIO À DECISÃO

2.1 Considerações Iniciais

Para que se possa compreender melhor as necessidades de novas ferramentas para aprimorar *web sites* de Grupos de Pesquisa e Desenvolvimento, serão abordados neste capítulo alguns tópicos relacionados aos conceitos básicos e à evolução das ferramentas para sistemas de apoio à decisão na *web.*

Segundo Power (2002), a concorrência acirrada e a maior exigência de qualidade por parte dos clientes estão forçando as organizações a se modernizarem, a serem mais criativas e mais eficientes na solução dos seus problemas. Para o autor, a tecnologia de informação está trazendo novos produtos e influenciando diretamente nos produtos fornecidos para o desenvolvimento das atividades nas organizações.

O alinhamento dos negócios com os sistemas de informação está sendo fundamental, uma vez que conhecer as tecnologias da informação que se transformam em ferramentas estratégicas é fator preponderante para o sucesso das organizações, pois fornece resultado para reestruturação dos *web* sites.

Segundo Dalfovo (2000a), entre as tecnologias da informação, uma nova categoria de Sistemas de Apoio à Decisão surgiu, conhecidos por *Business Intelligence*. Esses sistemas agregam soluções de *Data Warehouse, Data Mining* e *CRM (Customer Relationship Management*), que aborda conceitos de uma tecnologia relacionados a Clientes – Empresas e o *Data Warehouse* e *Data Mining* são tecnologias aos quais oferecem suporte ao monitoramento de iterações na *Web*.

Dessa forma, é preciso ter em mente que embora as vantagens que diferenciam o comércio eletrônico do comércio tradicional, são os relacionamentos, tornando mais eficiente a relação empresa – cliente, que precisa efetivamente ser utilizada de forma a agregar e dispor dos plenos benefícios que a tecnologia oferece, assim com o uso adequado dessas tecnologias o conceito pode ser ampliado para a relação usuários – fonte responsável pelo *site web*, uma vez que agora o gestor passa a gerenciar as interações que existem de seu *site* com os clientes.

2.2 Processos para a Tomada de Decisão

De acordo com Prates (1999), a decisão é uma das atividades que envolvem todos os indivíduos diariamente os quais, muitas vezes, não se dão conta de sua importância. O processo de tomada de decisão se alterou com a disponibilidade crescente de formação e tornou-se mais complexo. Hoje, existem mais fatores que influenciam no processo de tomada de decisão do que antigamente.

Para as empresas, houve uma mudança mais radical ainda nos seus conceitos de administração, porque com o advento da globalização e das Tecnologias de Informações, a competitividade ficou mais acirrada e os clientes mais exigentes. Também o seu processo decisório tornou-se muito mais complexo e o uso de ferramentas computacionais para dar suporte ao processo de tomada de decisão tornou-se indispensável, foi nesse sentido que surgiu a necessidade de se desenvolverem sistemas que proporcionassem o auxílio aos gerentes para que pudessem enfrentar os desafios da atualidade.

2.3 Sistemas de Apoio à Decisão

Sistemas de Apoio à Decisão (SAD) são sistemas baseados em computador, que auxilia o processo decisório, utilizando modelos para resolver problemas não estruturados. Os SADs analisam alternativas, propõem soluções, pesquisam o histórico das decisões tomadas, simula situações etc., participando diretamente do processo decisório (SPRAGUE, 1991).

Os Sistemas de Apoio à Decisão procuram ser mais flexíveis do que os Sistemas de Informações Gerenciais (SIG) e têm o potencial de auxiliar os tomadores de decisões em uma grande variedade de situações. Com o uso de Sistemas de Apoio à Decisão, os gestores poderão obter informações confiáveis, com valor agregado e disponível no tempo certo, agilizando o planejamento e a tomada de decisão nas empresas.

À medida que a informática vem evoluindo dentro das empresas (os sistemas operacionais já estão implantados, já existem sistemas que fornecem informações gerenciais, etc.), a tendência natural é que aumente a demanda por Sistema de Apoio à Decisão. O grande incentivo para a utilização de SAD se dará quando incorporarem a ele algumas importantes ferramentas computacionais utilizadas para o gerenciamento dos negócios: o *Data Warehouse, OLAP e o Data Mining*.

Segundo Rodrigues (1996), o fato de surgir uma nova geração de Sistemas de Apoio à Decisão não descarta e nem substitui os sistemas antigos e tradicionais. Muitas vezes, os novos sistemas trabalham em conjunto com os antigos para a solução dos problemas, para o gerenciamento dos negócios e para a elaboração de novas estratégias. Por exemplo, as informações obtidas por meio do *OLAP* ou do *Data Mining* podem alimentar um Sistema Multicritério de Apoio à Decisão ou qualquer outro sistema que trabalhe na linha de pesquisa operacional ou otimização. Os vários sistemas que foram desenvolvidos para realizar uma tarefa específica, continuarão a ter o seu lugar garantido e até mesmo continuarão a ser confeccionados e aperfeiçoados, principalmente os sistemas utilizados para otimização, como a maximização do uso dos recursos disponíveis.

Após uma melhor compreensão da evolução e dos fatores de influência do processo decisório, dos conceitos básicos, da evolução e das linhas de pesquisa dos Sistemas de Apoio à Decisão, é possível uma melhor assimilação em relação ao que será abordado sobre as três novas ferramentas de suporte à decisão. O próximo tópico abordará a primeira destas ferramentas: o *Data Warehouse* (PEREIRA e FONSECA, 1997).

2.3.1 Data Warehouse

Segundo Prates (1999), com a rápida evolução da tecnologia de informação e a disseminação do uso de computadores ligados entre si, praticamente, todas as organizações de médio e grande porte fazem algum uso de sistemas informatizados para realizar seus processos mais importantes. Com o passar do tempo, acaba-se gerando uma quantidade enorme de dados relacionados aos negócios, mas desintegrados entre si. Esses dados armazenados em uma ou mais plataformas funcionam como um recurso, mas, de modo geral, raramente servem como subsídio estratégico no seu estado original para o processo decisório.

Segundo Freitas *et al*., (2002a), os sistemas tradicionais de informática não são projetados para gerar e armazenar as informações estratégicas. Suas bases são formadas de dados cruciais à operação da organização. Em termos de decisão, os dados, de certa forma, são vazios e sem valor transparente para o processo gerencial das organizações. Essas decisões, normalmente, são tomadas com base na experiência dos administradores, podendo, também, ser fundamentadas em fatos históricos que foram armazenados pelos diversos sistemas de informação utilizados pelas organizações.

Para Domenico (2001), um *Data Warehouse* é projetado de forma que os dados possam ser armazenados e acessados de maneira a não ficarem restritos a Quadros e linhas relacionais. Como o *DW* está separado dos bancos de dados operacionais, as consultas dos usuários não impactam nesses sistemas, que desse modo ficam resguardados de alterações indevidas ou de perdas de dados. O *DW* contempla a base e os recursos necessários para um Sistema de Apoio à Decisão (SAD) e, principalmente, sistemas de Informações Executivas eficientes, fornecendo dados integrados e históricos que atendem desde a alta direção, que necessita de informações mais resumidas, até as gerências de baixo nível, em que os dados detalhados ajudam na observação de aspectos mais táticos da empresa. Nele, os executivos podem obter de modo imediato respostas para perguntas que normalmente não as possuem em seus sistemas operacionais e, com isso, tomarem decisões com base em fatos, e não apenas em intuições ou especulações.

Assim, um *DW* provê um banco de dados especializado, que gerencia o fluxo de informações baseando-se em banco de dados corporativos e fontes de dados externas à organização.

2.3.1.1 Conceitos Básicos

Para Inmon (1997), *Data Warehouse* é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para oferecer suporte ao processo gerencial de tomada de decisão.

Conforme Kondratiuk (1998), *Data Warehouse* é um processo em andamento aglutinador de dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas e *ad-hoc*, relatórios analíticos e de suporte à decisão.

Barquini (1996) define *Data Warehouse* como uma coleção de técnicas e tecnologias que juntas disponibilizam um enfoque pragmático e sistemático para tratar o problema do usuário final ao acessar informações que se encontram distribuídas em vários sistemas da organização.

Segundo Kimball *et al.,* (1998b), *Data Warehouse* é uma fonte de dados consultáveis da organização, formado pela união de todos os *Data Marts* correspondentes.

2.3.1.2 Elementos Básicos de um DW

Desde o seu surgimento, no início dos anos 90, a área de *Data Warehouse* tem perdido a precisão na definição de seus termos (Kimball, *et al*.,1998b). O Quadro 1 apresenta os conceitos dos principais elementos componentes de um *Data Warehouse*, segundo a visão integrada de (KIMBAL *et al*., 1998b).

Quadro 1 - Elementos Básicos de um Data Warehouse - Fonte Kimball et al., 1998b

ELEMENTOS BÁSICOS	DEFINIÇÃO
Sistemas de Origem	"Sistema operacional de registros cuja função é capturar as transações do negócio". p.14
Área de Estagiamento de dados	"Área de armazenamento e conjunto de processos que limpam, transformam, combinam, retiram duplicações, retêm, arquivam e preparam os dados fonte para uso no data warehouse" p.16 Servidor
Servidor de Apresentações	"Máquina física de destino onde estão armazenados e organizados os dados do data warehouse para consultas diretas dos usuários finais, dos geradores de relatórios e de outras aplicações" p.16.
Modelo Dimensional	"Disciplina específica para modelagem de dados que é uma alternativa ao modelo de entidades-relacionamento (model E/R)"p.17.
Processos do Negócio	"Conjunto coerente das atividades do negócio da organização, que fazem sentido aos usuários de negócio do data warehouse" p. 18.
Data Mart	"Um subconjunto lógico do data warehouse completo" p.18
Armazenamento de Dados Operacionais (ODS)	Ponto de integração com os sistemas operacionais da organização. Criados para integrar, em nível operacional, os diferentes sistemas da organização, sem, contudo, incluir consultas gerenciais, que ficam no nível do DW. p. 19-20.
OLAP	"Atividade genérica de consultar e apresentar dados textuais ou numéricos de data warehouses, bem como uma forma dimensional especifica de consultar e apresentar que é exemplificado por um número de 'vendedores OLAP'. Trata-se de uma tecnologia não-relacional e geralmente baseada em cubos multidimensionais de dados." p.21
ROLAP (OLAP Relacional)	"Conjunto de interfaces ao usuário e de aplicações que dão características multidimensional a bancos de dados relacionais". p. 21
MOLAP (OLAP Multidimensional)	"Conjunto de interfaces ao usuário, aplicações com base de dados proprietária que são fortemente multidimensionais" p. 21
Aplicação para Usuário Final	"Coleção de ferramentas que consultam, analisam e apresentam informações desejadas com vistas às necessidades do negócio da organização" p. 21
Ferramenta de Controle de Acesso aos Dados para Usuário Final	"Cliente do Data Warehouse (....). Uma ferramenta de controle de acesso aos dados para o usuário final. Pode ser simples como sistemas de consultas ad hoc ou complexas e sofisticadas com mineração de dados ou aplicações de modelagem". p. 21
Ferramentas de Consultas ad hoc	"Tipo específico de ferramenta de acesso dos dados que induz o usuário final a formar suas próprias consultas, manipulando diretamente Quadros relacionais e suas funções". p. 22
Aplicações de Modelagem	"Tipo sofisticado de ferramenta cliente do Data Warehouse com capacidades analíticas de transformar ou compreender as saídas do Data Warehouse" (e.g. Data Mining, modelos de previsão,modelos de comportamento, etc) p. 22
Metadados	"Toda informação no ambiente do Data Warehouse que não é dado real em si mesmo" p. 22

O trabalho de Kimball *et al*., (1998b) permite, apresentar uma visão geral dos elementos básicos de um *Data Warehouse*. A construção dessa tecnologia, em organizações universitárias, deve, também, refletir abordagens e soluções para cada um desses elementos.

Para Machado (2000), os sistemas de origem são formados pelos sistemas legados, geralmente, voltados para áreas como: pessoal, financeiro, estoques, etc. Devido à sua natureza operacional, tais sistemas são inadequados à formação de relatórios temporais e de difícil utilização pelos dirigentes das mais diversas instituições.

Na construção de elementos como área de estagiamento de dados, servidor de apresentações e armazenamento de dados operacionais (ODS), o projetista deve estar atento para a infra-estrutura disponível e, principalmente, para a arquitetura de informações da organização. Em instituições de ensino superior, a multiplicidade de plataformas, distanciamento físico das bases e heterogeneidade das aplicações acentuam a necessidade de cuidado redobrado na definição desses elementos. Um processo de negócio da organização envolve um conjunto de recursos de informações úteis e com um tema coerente com as atividades da instituição. Cada processo de negócio pode requerer um ou mais *Data Mart*. Em organizações universitárias, os processos de negócio podem ser identificados nas atividades meio e fim da instituição. De fato, dependendo da disponibilidade de seus sistemas legados, uma Instituição de Ensino Superior (IES) pode construir um *Data Marts* para suas áreas de compra, recursos humanos e gestão financeira ou para avaliação acadêmica, fluxo de egressos ou perfil pedagógico de seu quadro docente. Essas decisões também terão impacto na escolha dos demais elementos do *Data Warehouse*, em especial a opção pelo modelo *OLAP*, e das aplicações voltadas ao usuário final.

2.3.1.3 Modelo Estrela para o Data Warehouse

Conforme Valente (1996), modelos de bases de dados relacionais apresentam tabelas com relacionamentos complexos e com múltiplas uniões circulares entre dois pontos do modelo. Para a maioria dos usuários que utiliza ferramentas para compor suas consultas, é necessário que o acesso à base de dados seja simples o suficiente para facilitar o sucesso nas operações. Para acomodar as necessidades de todos os usuários e facilitar a atualização do *DW* o projetista deve criar um modelo, cujo usuário final possa entendê-lo, facilmente, em termos do negócio.

Segundo Freitas *et al*., (2002a), o principal tipo de modelo dimensional é o chamado modelo Star (Estrela), no qual existe uma tabela dominante no centro, chamada de tabela de fatos, com múltiplas junções, conectando-a a outras tabelas, sendo essas chamadas tabelas de dimensão. Cada uma das tabelas secundárias tem apenas uma junção com a tabela central. O modelo Estrela, representado na Figura 3, tem a vantagem de ser simples e intuitivo, e também por utilizar enfoques de indexação e união de tabelas. Para Kimball (1996b), o modelo entidade-relacionamento não é o mais adequado para a análise dos dados no ambiente gerencial. O modelo dimensional é o mais apropriado para esse ambiente. A Figura 3, a seguir apresenta o modelo dimensional schema estrela.

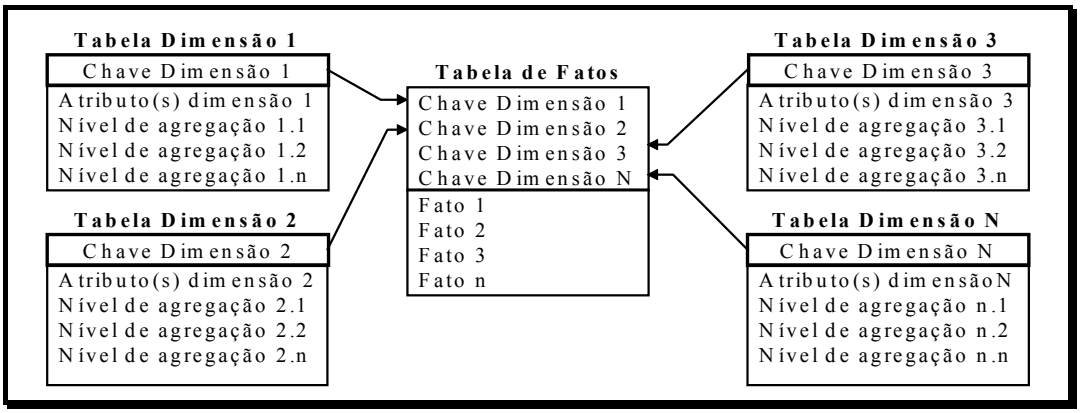


FIGURA 3 – MODELO SCHEMA ESTRELA - FONTE FREITAS *ET AL.*,(2002)

Segundo Domenico (2001), a tabela de fatos contém milhares (ou milhões) de valores e medidas do negócio da empresa, como transações de vendas ou compras. Cada uma dessas medidas é tomada segundo a interseção de todas as dimensões. Uma característica importante da tabela de fatos é a esparsidade, ou seja, se não existe um cruzamento para alguns valores das dimensões, por isso a tabela de fatos não armazena zeros.

As tabelas de dimensão armazenam as descrições textuais das dimensões do negócio. Cada uma dessas descrições textuais permite definir um componente da respectiva dimensão. Uma das principais funções dos atributos de tabelas de dimensão é servir como fonte para restrições em uma consulta ou como cabeçalhos de linha no conjunto de respostas do usuário. Tabelas de dimensões tendem a utilizar tipos de caracteres em vez de numéricos, de forma que suas linhas são muito mais longas, mas em pouca quantidade, ocupando pequena porcentagem de espaço em disco. As tabelas de fatos podem utilizar até 95% da área destinada ao *Data Warehouse* (BARQUINI, 1996).

Um fator importante relacionado à tabela de fatos é que, como representa os relacionamentos muitos-para-muitos entre as tabelas de dimensão, esta tem como chave primária uma chave composta de todas as chaves estrangeiras das tabelas de dimensão (KIMBALL, 1996b).

Para um bom desempenho do modelo Estrela, é necessário que os projetistas saibam antecipar, na modelagem do DW, as consultas mais freqüentes a serem realizadas pelos usuários.

Segundo Kimball (1996b), o modelo dimensional apresenta várias vantagens no que diz respeito à sua utilização para o *DW*, dentre essas estão:

- arquitetura padrão e previsível;
- dimensões do modelo são equivalentes, ou seja, podem ser vistas como pontos simétricos para a Quadro de fatos;
- modelo dimensional é flexível, pois permite inclusão de novos elementos de dados, bem como mudanças ocorridas no projeto;
- facilidade na alteração das Quadros de fatos e dimensão;
- todas as aplicações existentes anteriormente à mudança permanecem rodando sem problemas.

2.3.1.4 As Ferramentas Utilizadas em um Data Warehouse

Segundo Figueiredo (1998), existem várias ferramentas utilizadas em um *Data Warehouse*:

- ferramenta para Armazenamento: são os banco de dados, considerados o coração do *Data Warehouse* e parte imprescindível do projeto;
- ferramenta para a extração de dados: busca na base de dados operacionais os dados que serão armazenados no *Data Warehouse;*
- ferramenta para a transformação de dados: ajusta os dados para o formato do *Data Warehouse.* Esse formato auxilia as futuras pesquisas;
- ferramenta para a limpeza de dados: efetua os ajustes necessários nos dados, fazendo correções, desmembramento e fusões de dados, quando necessário, visando melhorá-los para facilitar as futuras pesquisas;

- repositórios de metadados: estão intimamente relacionados às ferramentas de extração. Metadados são as definições dos dados que permitem saber a origem da informação, bem como as vezes em que ela foi alterada. Sua função é manter a consistência dos dados;
- transferência de dados e replicação: pode ser considerado um subconjunto das ferramentas de extração. Não faz nenhum tipo de processamento e transformação, apenas transfere um dado de um lugar "A" para "B". Geralmente, é utilizado para facilitar e dar uma resposta mais rápida às consultas ou análises, movendo os dados para um lugar apropriado e fazendo o que for necessário para agilizar o serviço solicitado;
- gerenciamento e administração: é a típica ferramenta que só faz sentido depois que o *Data Warehouse* está construído. Monitora o dia-a-dia, como a performance e segurança do sistema;
- *query* ou ferramentas para gerenciamento de consultas: fazem consultas ou geram relatórios retirando os dados do *Data Warehouse*, resumindo-os e apresentando-os em um formato apropriado;
- ferramentas para gerenciamento de relatórios: são semelhantes às ferramentas do item anterior, porém elas estão voltadas para a geração de relatórios mais complexos contendo, por exemplo, relatórios sintéticos e analíticos em conjunto, gráficos e outros tipos de visualização dos dados;
- ferramentas de *OLAP*: é a parte mais visível do *Data Warehouse* porque é por meio dessas ferramentas que se procede à análise dos dados. Auxiliam os gerentes a sintetizarem as informações sobre a empresa por meio de comparações, visões personalizadas, análise histórica e projeção de dados;
- *Data Mining* para Nimer e Spandri (1998), é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em *Data Warehouse* usando técnicas de reconhecimento de padrões, estatísticas e matemáticas;
- simulação: projetam cenários respondendo a perguntas do tipo "e se", por exemplo: “e se os juros aumentarem, qual será o comportamento de minhas vendas”.

2.3.2 Mineração de Dados

Segundo Nimer e Spandri (1998), os avanços da chamada "era da informação" têm colocado como desafio à implementação de técnicas capazes de mensurar e descobrir padrões relevantes na crescente massa de dados, resultante, sobretudo, do aumento da complexidade das tarefas operacionais e decisórias. Essa explosão nos dados pode determinar a sobrevivência ou não de uma organização, desde que esta consiga extrair informações úteis à tomada de decisão e à melhoria nos processos operacionais.

Essa visão da análise de dados baseia-se em duas novas tendências: na avalanche de informações e no questionamento sobre esses dados (Hair *et al.,* 1998). Nesse enfoque, uma nova perspectiva é apresentada, na qual a análise de dados é vista com um caráter mais exploratório.

De maneira mais abrangente, encontra-se a *Knowledge Discovery* in Database (KDD), sendo esta a designação para o processo que envolve a seleção, o pré-processamento e a transformação dos dados, bem como a aplicação de algoritmos, a interpretação dos resultados e a geração de conhecimento, como ilustrado na Figura 4.

Segundo Fayyad (1996a), no interior desse processo encontra-se o *Data Mining*, ou Mineração de Dados (MD), que é uma etapa no processo de KDD, responsável pela aplicação dos algoritmos, cuja finalidade é a identificação de padrões Ej sobre uma base de dados F a geração de um conjunto de regras descritivas do comportamento de uma base de dados.

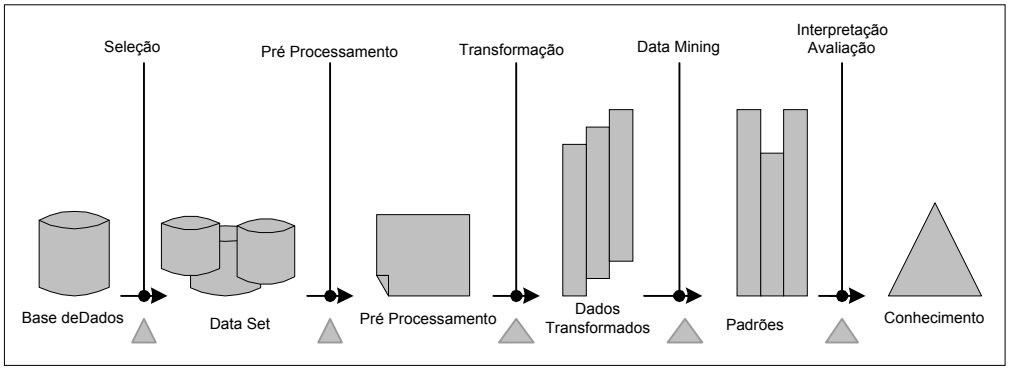


FIGURA 4 - PROCESSO DE KDD - FONTE ADAPTADO DE FAYYAD *ET AL*, (1996B)

A proposta de MD é proporcionar uma perspectiva nova ou, mais precisamente, uma evolução nos processos de análise, permitindo a descoberta de novos padrões ou a validação de padrões conhecidos. Tais análises são geralmente efetuadas em grandes quantidades de dados.

Por outro lado, para Harrison (1998), MD é a exploração e análise, por meios automáticos ou semi-automáticos, de grandes quantidades de dados, cuja finalidade é a descoberta de modelos e regras significativas.

Na visão de Berry e Linoff (1997), existem duas metas primárias para um sistema de mineração de dados:

- previsão: envolve a utilização de algumas variáveis (atributos da base de dados) para prever valores desconhecidos ou futuros de outras variáveis de interesse;
- descrição: procura por padrões que descrevam os dados e que sejam interpretáveis.

Levando-se em consideração essas metas, pode-se observar a real importância desse processo dentro de uma organização, onde tais metas devem prever algumas fases. Harrison (1998) identifica quatro fases:

- identificar problemas e áreas em que a análise de dados pode fornecer valor;
- transformar dados em informações acionáveis, usando técnicas de mineração de dados;
- agir sobre a informação e, com base nela, melhorar os processos que regem o relacionamento da empresa com seus consumidores e fornecedores;
- medir os resultados dos esforços para fornecer idéias sobre como explorar os dados. Essa fase proporciona o *feedback* para o aumento constante na qualidade dos resultados.

O processo de descoberta de conhecimento envolve algumas etapas, entre elas a definição do domínio da aplicação, a limpeza e o pré-processamento dos dados, a representação dos dados, a mineração de dados e a interpretação dos resultados.

Primeiramente, deve-se definir o problema e as metas desejadas pelo usuário. Nesta fase, devem preocupar-se com critérios de desempenho no domínio da aplicação e a interoperabilidade com o usuário final.

Na etapa seguinte, são realizados o pré-processamento e a limpeza dos dados pela remoção de ruídos ou dados inválidos que atrapalhem o processamento, bem como a adoção de estratégias para manusear campos que apresentem dados perdidos.

Em seguida, a representação dos dados procura modelá-los de maneira que eles possam ser utilizados por algum algoritmo de extração de conhecimento. Como exemplo,

pode-se citar a transformação de valores lingüísticos em valores numéricos dentro de um domínio ou a transformação de valores contínuos para discretos.

A etapa de mineração de dados propriamente dita constitui na busca, em uma base de dados, por informações relevantes, de difícil identificação. De acordo com Fayyad (1996a), a busca é realizada em três etapas: primeiramente, decide-se se o processo será de classificação, agrupamento ou sumarização; em seguida, escolhe-se um dos métodos a serem utilizados na busca por padrões; e, por último, efetua-se o processo de busca ou a mineração dos dados. Uma grande variedade de técnicas analíticas tem sido utilizada em mineração de dados. Essas técnicas vão desde as tradicionais estatísticas multivariadas, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos. O Quadro 2 demonstra as principais funções da mineração de dados, algoritmos utilizados e exemplos de aplicações. A escolha de uma ou de outra função depende essencialmente do negócio, da aplicação e da quantidade e qualidade dos dados disponíveis.

Quadro 2 - Funções da Mineração de Dados (Bigus, 1996)

Funções	Algoritmos	Aplicações
Associação	Estatística, teoria dos conjuntos	Análise de mercados
Classificação	Árvores de decisão, redes neurais, algoritmos genéticos	Controle de qualidade, avaliação de riscos
Agrupamento	Redes neurais, estatística	Segmentação de mercado
Modelagem	Regressão linear e não linear, redes neurais.	*Ranking* de clientes, controle de processos, modelos de preços
Previsão de séries temporais	Estatística, redes neurais	Previsão de vendas, controle de inventário
Padrões Seqüenciais	Estatística, teoria dos conjuntos	Análise de mercado sobre o tempo

Por último, a etapa de interpretação dos dados verifica a validação do conhecimento extraído da base de dados, apresentando esse conhecimento de maneira mais simplificada utilizando gráficos, Quadros e regras. O conhecimento extraído pode ser validado por meio de métodos estatísticos ou pelo parecer de um especialista.

2.3.3 Customer Relationship Management (*CRM*)

Este tópico apresenta de forma bem clara os mais diversos conceitos de *Customer Relationship Management* (CRM) para o contexto dos negócios, os fatores-chave de sucesso e

a forma da evolução dos tradicionais *Call Centers* para o Contact Center. Essa seção vai explorar a forma e a maneira com que diversos autores abordam o assunto.

Segundo Makenna (1991), a idéia básica do CRM, consiste em fazer com que, baseando-se em análises das informações geradas nos contatos, as transações entre o cliente e a empresa se transformem em relacionamentos duradouros, tornando-o fiel a determinado produto e serviço. Nesse processo, Mcclain (2000) diz que são utilizados todos os conhecimentos adquiridos em qualquer contato com o cliente. É uma forma de análise do comportamento dos clientes e, considerando as lições tiradas dessa análise, a maneira de influenciar o seu comportamento, antecipam-se as suas necessidades; ou ainda, um conjunto de políticas práticas e infra-estrutura tecnológica, visando a reter o cliente por meio da excelência no atendimento. Dessa forma, o CRM é um processo contínuo, complexo, que tem começo, mas não tem fim, ou seja, o CRM é uma estratégia para se aproximar do cliente; um marketing específico para esse relacionamento.

Bretzke (2001a) destaca que o marketing em tempo real e as ações de decisões referentes aos clientes atuais que precisam estar fortemente alicerçadas em informações que agilizem e otimizem todo o processo de vendas e de atendimento. As informações de relacionamento precisam ser compiladas ou recuperadas no momento em que o contato entre a empresa e o cliente está ocorrendo, para que se possa conhecer e reconhecer o cliente e, conseqüentemente, direcionar produtos, serviços e ofertas completamente ajustadas a ele, que, por conseguinte, estará disposto a estabelecer a preferência pela marca, repetir a compra e inclusive pagar mais para obter o valor agregado que lhe é oferecido.

Para Stone (2001), CRM é um dos métodos mais sofisticados e eficientes que transformam a maneira como as empresas podem aumentar a rentabilidade dos clientes atuais. Além disso, o uso da Internet como canal de relacionamento de vendas é amplamente facilitado e viabilizado por esse novo método, que, embora seja praticado por poucas empresas, tem mostrado resultados largamente compensadores para os clientes mais leais, como a maior satisfação com a marca e um nível de proximidade nunca antes experimentado.

Nesse contexto, o *Call Center* transforma-se num *Contact Center* gerenciando todo e qualquer contato do cliente com a empresa, por meio da Internet, do fax ou do telefone, respondendo em tempo real a qualquer solicitação ou pedido de compras.

Customer Relationship Management ou Gerenciamento do Relacionamento com o Cliente, como o próprio nome indica, é a integração entre o Marketing e a Tecnologia da Informação, cujo objetivo é prover a empresa de meios mais eficazes e integrados para atender, reconhecer e cuidar do cliente, em tempo real, e transformar seus dados em informações que, disseminadas pela organização, permitem ao cliente ser "conhecido" e cuidado por todos e não só pelas operadoras do *Call Center* (LIEB, 1999).

Segundo Bretzke (2000), o CRM é a combinação da filosofia do Marketing de Relacionamento, que ensina a importância de cultivar os clientes e de estabelecer com ele um relacionamento estável e duradouro por meio do uso intensivo da informação, com a Tecnologia da Informação, que provê os recursos de informática e telecomunicações integrados de uma forma singular que transcende as possibilidades dos *Call Centers* atuais. O marketing personalizado e o atendimento diferenciado destacarão a empresa das demais concorrentes. Essa estratégia permite à empresa realizar previsões de vendas, gerenciamento de clientes e de fornecedores, buscando sempre o seu cliente fiel e lucrativo.

Com base nessas definições, é possível que o conceito fundamental de *CRM* esteja bem difundido no mercado, ou seja, o *CRM* se tornará importante nesse novo ambiente de negócios não somente como solução, mas também como uma metodologia fundamentada em seus conceitos próprios.

Mais do que isso, o *CRM* terá um papel de unificação de um conjunto de áreas baseadas em tecnologia, compondo uma nova família, uma nova geração, focada no atendimento aos melhores clientes de uma organização, trazendo-os novamente em busca dos bens e serviços oferecidos.

Essa integração singular pressupõe predisposição da empresa na manutenção de um relacionamento suportado por processos operacionais mais ágeis e a seleção de tecnologia adequada, o que requer metodologia com experiência comprovada nesse tipo de solução. Isso significa que essa interação é uma grande virada no conceito de atendimento ao cliente, porque extrapola a prática existente em qualidade e a possibilidade de aumentar a fidelidade do cliente, e, conseqüentemente, a rentabilidade da empresa.

Segundo Rapp (1996) e Bretzke (2001a), Argumentam que os recursos humanos precisam ser treinados e capacitados, em todos os níveis, não só para melhorar a qualidade do atendimento, mas também para usar adequadamente as informações transformadoras de

possibilidades de negócios em lucros. Para os autores citados, as vantagens de operar num ambiente de *CRM* integrado são:

- redução de custos da operação de vendas;
- geração de novos negócios e oportunidades de receitas;
- integração com um ambiente multimídia como a Internet, por exemplo.

Nesse contexto, existe um bom número de soluções à disposição, entretanto há também muitas dúvidas em torno do Marketing de Relacionamento (*CRM*), ou seja, questiona-se se a sua tecnologia é mais apropriada para as soluções.

Finalmente, Bretzke (2001c) recomenda a contratação de profissionais experientes em metodologia de implantação de projetos, para que façam o papel de integradores, que atuem em conjunto com a área de Tecnologia da Informação e Marketing, e que conciliem os interesses das áreas sob a égide da filosofia de Marketing de Relacionamento. Para tanto, as empresas integradoras do *Marketing* e da Tecnologia da Informação, como esforço de manterem sua posição competitiva, estão concentrando-se para oferecer cada vez mais serviços aos seus clientes, pois entendem que a lealdade dos clientes diminui a sua dependência da inovação de produtos e serviços, tornando-as menos suscetíveis à guerra de preços e colocam o diferencial competitivo na lealdade do cliente, investindo em *Call Centers*.

Portanto, adotar o quanto antes o método do *CRM* é uma questão de manter a competitividade, pois os clientes estão constantemente voltando a atenção para empresas que aumentam as suas expectativas e não se contentando simplesmente com um acesso rápido e fácil, a qualquer hora, às centrais de atendimento.

2.4 Considerações Finais

Este capítulo demonstrou a evolução dos fundamentos teóricos quanto aos Sistemas de Apoio à Decisão, compreendendo as mudanças sofridas por esses processos com o surgimento dos Sistemas de Informações. As Tecnologias de Informações estão provocando mudanças e alterações na organização do processo de trabalho e o sucesso das organizações, passa pela modernização e atualização das novas tecnologias.

Compreender a finalidade dos Sistemas de Apoio à Decisão, configurada a fundamentação do *Data Webhousing*, tem sido um desafio à estruturação dos *web sites* de grupos de P&D. Contudo, procurou-se demonstrar o desenvolvimento das três principais ferramentas de apoio à decisão: o *Data Warehouse*, o *Data Mining* e *o CRM.*

Dessa forma, além de questões intrínsecas a *web*, há questões relativas ao nível de exigência dos usuários. Com a evolução dos *web sites*, entretanto, a análise de *Clickstream* tornou-se uma ferramenta importante para a construção de *web site*, em que as interações com os usuários são, na maioria das vezes, completamente virtual, e, por isso as novas tecnologias apresentadas devem acompanhar essa virtualização das interações.

Este capítulo teve o objetivo de apresentar subsídios para que se possa entender melhor o próximo capítulo que apresentará uma revisão teórica de *Data Webhousing* e a viabilidade de uma proposta metodológica de construção de uma aplicação *Data Mart Clickstream*. Considerando-se essa aplicação e o resultado das técnicas aplicadas, será proposta melhoria para *site* de grupos de P&D no capítulo final desse trabalho.

3 DATA WEBHOUSING

3.1 Considerações Iniciais

A necessidade das organizações obterem informações úteis para tomada de decisão em seus bancos de dados fez com que emergisse uma tecnologia chamada *Data Warehous*e.

Segundo Silberschatz (1999), *Data Warehouse* é um repositório de informações coletadas em diversas fontes, armazenadas sob um esquema único, em um só local. Uma vez coletados, os dados são armazenados por um período longo, permitindo acesso a dados históricos.

De acordo com Oliveira (1998), *Data Mart* é um *Data Warehouse* de menor porte, construído para armazenar dados ligados a determinado aspecto do negócio da organização. As diferenças entre ambos são:

- volume de Informações: *Data Mart* é criado para localizar informações necessárias para uma unidade ou função específica da organização. Já um *Data Warehouse* é construído para oferecer informações necessárias a toda a organização;
- tratamento da Informação: *Data Warehouse* gerencia grandes quantidades de informações. *Data Mart* está focado primeiramente no sumário ou exemplos de informações;
- gerenciamento: *Data Warehous*e por suas características é gerenciado pelos estrategistas da organização, enquanto que o *Data Mart* pode ser gerenciado por gerentes ou encarregados de setor.

Dessa forma então, tornou-se necessária a criação de um *Data Warehouse*, ou seja, um *Data Mart* focado nos usuários dos *sites da web*, ao contrário dos *Data Warehouses* tradicionais, que armazenam informações referentes ao desempenho da organização e sumarizando dados para definição de estratégias competitivas. A união entre *web* e *Data Warehouse* origina o chamado *Data Webhous*e (KIMBALL, 2000f).

Nesse caso, segundo Kimball e Merz (2000f), quando se fala na atividade de utilizar o *Data Warehouse* para a w*eb* se constitui em disponibilizar todos os serviços dos Sistemas Operacionais para a *web.* Já em relação à atividade de utilizar a *web* para o *Data Warehouse* é

definido como armazenamento dos dados da seqüências de clique *(clickstrea*m). A Figura 5, apresenta um armazém de dados e de seqüências de clique.

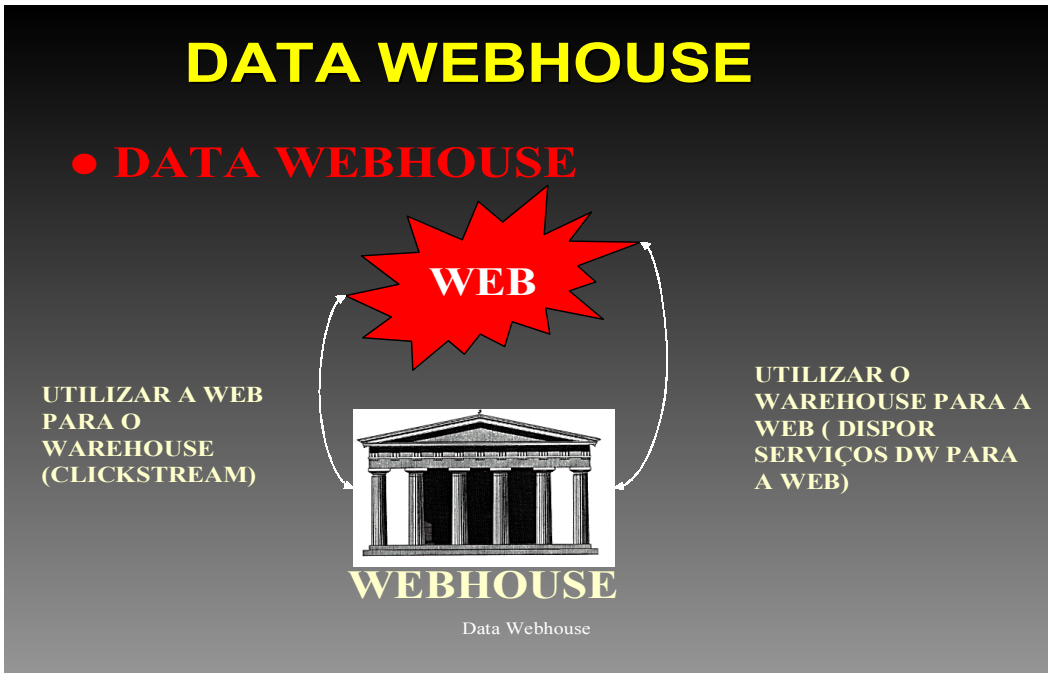


FIGURA 5 – DATA WEBHOUSE – FONTE GONÇALVES *ET AL*., (2001)

Para Ribeiro (2001), a revolução da *web* não diminuiu a importância do *Data Warehouse*, ao contrário, aumentou e em muito, a expectativa das pessoas a respeito dos tipos de informação que poderiam ser disponibilizados via *web*. Dessa forma, surgiu a união entre o *Data Warehouse* e a *web*, na qual os responsáveis pelo controle e gerenciamento das empresas poderem utilizar interfaces w*eb* para acessar o *Data Warehouse*.

De uma outra forma, o *Data Warehouse*, na sua missão de armazenar informações vitais para o processo de decisão, passou a ser útil para armazenar e monitorar a interação dos clientes com o *web site* da empresa. Com a adoção de *web sites* pelas empresas e o crescimento do uso da *web* por usuários, tanto domésticos quanto corporativos, o empreendedor que possuir maior conhecimento sobre seu usuário na *web* terá um diferencial competitivo em relação aos seus concorrentes.

Para Kimball (2000f), o *Data Webhouse* permite analisar todo o caminho realizado por um visitante em um *site da web*, mapear cada clique, e conhecer melhor cada usuário por meio dos dados mantidos nos arquivos de *log* do servidor, tornando, portanto, muito mais fácil e real uma análise de cada evento realizado no interior do *site* pelo visitante.

Como o objetivo do *Data Webhouse (DW)* é capturar, armazenar e publicar dados de seqüências de clique, a organização que possuir informações dos cliques de seus clientes, terá grande vantagem competitiva sobre seus concorrentes, uma vez que, de posse desses dados e da evolução da tecnologia DW, a organização poderá descobrir informações valiosas, como as que seguem abaixo:

- o local mais visitado do seu *site;*
- o local que apresenta maior número de vendas no seu *site;*
- o local menos visitado no seu *site;*
- a página do seu *site* é vista como sessão final, onde os usuários geralmente encerram a sessão;
- o ponto em que o novo usuário clica nas primeiras visitas de acordo com o seu perfil;
- o perfil de navegação de um cliente existente;
- o Navegador mais utilizado no seu *site;*
- o documento mais visado dentro do *site;*
- a página com maior índice de entrada e saída no *site;*
- os arquivos mais copiados do *site.*

Segundo Kimball e Merz (2000f), para se ter êxito com essas informações, o *site* deve ser bem projetado, uma vez que incorporará ferramentas que monitorarão os seus usuários e os seus visitantes.

Com a introdução da seqüência de clique como fonte de dados, torna-se necessário trabalhar no processo de análise de dados capturados, desde a coleta dos dados até o seu armazenamento.

Assim, este estudo mostrará, igualmente os processos e as fases de preparação para o desenvolvimento de um projeto de *Data Webhouse,* tomando por base as metodologias propostas por Kimball (2000f), Barbosa et al. (2002) e Farias (2002).

3.2 Gerenciando as Principais Característica do Data Webhouse

Segundo Kimball (2000), ao se analisar a *web*, observa-se que trata de uma imensa fonte de dados relativos ao comportamento de indivíduos, quando estes interagem com *web sites* por meio de um *browsers*. Apesar de esses dados estarem desorganizados e sem nenhuma forma de tratamento, eles têm potencial para prover detalhes, até agora desconhecidos, sobre a utilização de *web sites* por qualquer ser humano. Esses dados são denominados *Clickstream*, ou seja, seqüências de cliques.

O *Clickstream* é considerado uma imensa fonte de dados não disciplinados, que é trazida ao ambiente de *Data Webhouse* para análise, seja por si própria, seja por combinação com fontes de dados convencionais. Embora seja imperfeito, o registro da história de navegação permite a realização de análises que anteriormente não poderiam ser realizadas.

Se antes podiam registrar-se apenas transações efetivamente realizadas entre o usuário e a empresa, tal como compra, encomenda e devolução, o histórico da navegação do usuário cria uma possibilidade de captura de intenção do usuário, porque parte do "caminho mental" percorrido pelo usuário, entre o portal e a *home page*, e da efetivação de algumas transações que ficam registradas nos arquivos de logs dos servidores de páginas. Com a possibilidade de integração desses registros de navegação com o *Data Webhouse* corporativo, será possível fazer análises da intenção do usuário na sua efetiva ação (KIMBALL, 2001c).

3.2.1 Utilizar a Web para o Webhouse

Segundo Gonçalves e Oliveira (2001), o trabalho de trazer a *web* para o *Data Warehouse,* formando um *Data Webhouse*, significa mudar o comportamento do *Data Warehouse.* Originalmente, o *Data Warehouse* possui informações cuja extração é feita nos sistemas transacionais. O trabalho de capturar informações das transações já é largamente conhecido, a novidade é justamente como capturar, analisar e entender o comportamento dos cliques dos clientes nos *web sites.*

De acordo com Voelcker (2001), o acesso diário aos *sites da web* permite tratar de uma imensa fonte de dados referentes ao comportamento dos usuários da Internet, na qual por meio da captura, análise e entendimento do comportamento dos clique dos usuários nos *sites*

da web, permite descobrir informações preciosas a respeitos dos usuários. Esta abordagem é conhecida como seqüência de cliques ou *Clickstrea*m.

O Clickstream é, literalmente, um registro de todas as ações feitas por qualquer visitante a um *web site*, por isso, essas fontes de dados têm um potencial muito maior que as fontes de dados tradicionais (KIMBALL, 2000f), pois a seqüência de clique, apresenta informações do comportamento das pessoas na *web*.

Segundo Inmon (2001), o *Clickstream* tem como objetivo suprir as deficiências das fontes de dados tradicionais no ambiente *web*. A seqüência de clique não é somente mais uma fonte de dados que foi extraída, limpa e organizada no *Data Warehouse*. Ela é, na verdade, uma coleção de fontes de dados, já que existem diversas formas de registrar o comportamento dos usuários, e de acordo com suas necessidades, podem identificá-los nas sessões mais visitadas no *site*.

3.2.2 Tipos de Aplicação e de Análise

Conforme Schonberg (2001), o "*Clickstream"* são dados capturados dos usuários através da navegação em *web sites* que pode transformar-se em informações relevante para a tomada de decisões nas áreas de: Marketing, Merchandising, Comunicação, Infra-estrutura de serviços, e também nas áreas de negócios da empresa.

O Marketing na *web* torna-se mais efetivo, porque pode ser feito com base nas características individuais do cliente. Marketing externo que é publicado em outros *web sites*, também se torna mais efetivo, pois a possibilidade de identificação das referências externas (de outras páginas para o *web site* da empresa) no *log* do servidor de páginas aliadas à informação das estatísticas de acesso às páginas, fornece uma medida da eficiência das campanhas de Marketing.

Schonberg (2001), ressalta ainda que o *Clickstream* pode auxiliar na descoberta de informações para o Merchandising *online* que trata da aquisição e disposição dos bens e serviços no *site* virtual, de forma a atender alguns objetivos de negócio, por exemplo, o estímulo às compras. Dessa forma será possível então realizar outros tipos de análise nessa área, pois o Merchandising possui as mesmas métricas do Marketing, que são indicadores de retorno financeiro para a empresa.

Gonçalves (2001) destaca que, na área de comunicação, uma das principais vantagens do *Clickstream* é a melhoria na customização do conteúdo. Por meio da identificação do usuário na *web* e do relacionamento da informação com os registros históricos, transacionais, demográficos e do perfil do usuário, é possível identificar os conteúdos que mais interessam para cada um deles. Outra conseqüência para a comunicação é a resposta sobre a efetividade do *web site* ser mais versátil.

3.3 Monitorando as Ações dos Usuários de um Web Site

De acordo com Kimball (2000f), cada *site* bem projetado incorpora ferramentas para monitorar os *web sites*, seus visitantes e suas ações, como um meio de medir a eficácia e o impacto do *site.* Essas informações de monitoramento podem ser então, utilizadas para personalizar o conteúdo dinâmico do *site*, aumentando a relevância e o interesse do usuário, à medida que a sessão progride ou quando ele retorna ao *site* mais tarde.

Dessa forma, existem duas formas de coleta de dados. A primeira é chamada de seqüência de clique, que é registrada pelo mecanismo de *log* do servidor da *web.* A segunda é fornecida por servidores de aplicativos do *site,* sendo composta pelos dados que seriam capturados por qualquer aplicativo, tais como: entrada de pedidos, pesquisa de texto ou relatório de crédito. No *Data Webhouse,* mesclam-se essas fontes de informações em uma única forma coerente.

Para Kimball (2001), as ações do usuário na *web* podem ser motivadas por um número de necessidades diferentes, as quais podem mudar de um momento para outro, durante a sessão do navegador. A seguir, estão algumas das ações mais comuns que um usuário pode realizar durante uma sessão da *web:*

- pesquisa: consiste em localizar um produto, serviço ou fonte de informações específicas;
- coleta de informações: comparar produtos e serviços, ler FAQs, diversão;
- educação: utilizar manuais, classes interativas, livros on-line e artigos;
- comunicação: participar de grupos de discussão e de notícias, utilizar correio eletrônico;
- fazer download;

- compras e pedidos: selecionar e comprar produtos, freqüentemente, com cartão de crédito;
- entrada acidental: clicar no botão ou objeto errado, erros de URL, links quebrados.

A cada uma dessas ações, o usuário pode ter necessidades especiais de banco de dados, tanto para monitorar suas ações, quanto para entregar conteúdo dinamicamente em uma situação, por exemplo, pesquisar um *site na web,* pode requerer um sistema de pesquisa de texto. (KIMBALL, 2000f).

3.3.1 As Técnicas de Monitoração

Para Kimball (2000e), existem várias perguntas importantes para a área de marketing e para os *webmasters* que determinam a eficácia da promoção do *site*, entre elas: de onde veio o visitante? Como ele encontrou o *site*? Como chegou a uma página (imagem ou formulário) específica?

A identificação de uma sessão (ID de sessão) apresenta os registros de cada ação individual do usuário em uma sessão, quer sejam derivado das seqüências de cliques, quer da interação com aplicativos que deverão conter uma *tag* especial de identificação própria e exclusiva.

Em muitos casos, um ID de sessão poderá não estar disponível imediatamente, quando os eventos iniciais relacionados à sessão estiverem sendo registrados. Nesses casos, um ID temporário de sessão será necessário e, mais tarde, a situação será resolvida com um ID de sessão aceitável pela empresa, que seguirá as informações de log por meio do *Data Warehouse.*

Kimball (2000) ressalta que apesar dos protocolos http não oferecerem informações de estado (ou seja, falta o conceito de sessão), existem várias formas de solucionar esse problema. Um método é permitir ao servidor colocar um *cookie*[3] de nível de sessão no navegador do usuário. O valor do *cookie* pode servir como um ID temporário de sessão não apenas para o navegador, mas também para qualquer aplicativo que solicite o *cookie* de sessão

[3]*Cookie* - é uma informação que um servidor *web* pode armazenar temporariamente junto a um *browser.*

ao navegador. Essa solicitação deve vir do mesmo servidor da *web* que colocou o *cookie* em primeiro lugar (KIMBALL, 2000e).

A *Secure Sockets Layer* (*SSL*) do HTTP oferece uma oportunidade de monitorar uma sessão do usuário, porque pode incluir uma ação de login pelo usuário e pela troca de chaves de criptografia. Se a geração da página é dinâmica, pode-se manter o estado do usuário colocando um ID de sessão em um campo oculto de cada página retornada ao usuário. Além disso, o *site da web* pode estabelecer um *cookie* persistente na máquina do usuário que não é excluído pelo navegador quando a sessão termina. (KIMBALL, 2000e).

Conforme Kimball (2000f), o método mais confiável de monitoramento de sessão de registros de *log* do servidor *web* é obtido pela configuração de um *cookie* persistente no navegador do usuário. Resultados menos confiáveis, porém bons, podem ser obtidos pela configuração de um *cookie* de nível de sessão não-persistente e pela associação de entrada de log contíguas no tempo do mesmo *host.* O último método requer um algoritmo robusto no pós-processamento de *log* para assegurar resultados satisfatórios e para decidir quando não levar os resultados a sério.

3.3.2 Análise Comportamental

Para Mena (1999), o comportamento do usuário durante uma visita a um *site da web* pode fornecer "*insights*" valiosos sobre a eficácia do *site*, bem como sobre os hábitos de navegação do usuário como mostram os tópicos abaixo:

- ponto de entrada: muitos usuários entrarão em um *site* por meio da sua *home page,* simplesmente porque digitaram a URL do *site* em seu navegador. Eles também podem entrar por meio de um link em outro *site.* As informações de ponto de entrada são importantes para marketing e projeto, porque toda página comumente utilizada para entrada deve convidar o usuário a explorar seu site por inteiro;
- permanência: a permanência é o tempo em que o usuário, realmente, tem uma página visível no navegador. Capturar essa informação para um *site da web* é o mesmo que ser capaz de assistir a alguém lendo uma revista e medir com o cronômetro o tempo que tal pessoa gasta para ler cada página. Se o tempo de permanência em uma página for muito curto, pode-se supor que a página foi acionada erroneamente ou que seu conteúdo é irrelevante para o usuário;

- consulta: os argumentos de pesquisa que um usuário digita em um formulário da *Web* podem dizer muito sobre as indicações dos usuários e a usabilidade do *site*. Para monitorar uma pesquisa de forma livre, é preciso capturar as palavras-chave, bem como as contagens dos resultados e o próprio resultado;
- navegação *intra-site*: a maneira pela qual um usuário navega por um *site da web* pode fornecer padrões de medida de valor adicional para os projetistas do *site.* O estilo de navegação de um usuário deriva da seqüência de clique que pode ser utilizada para ajustar o *site* e para otimizar apresentações futuras de informações utilizando o gerenciamento de conteúdo dinâmico;
- ponto de saída: quando o usuário sai do *site*, ele geralmente não deixa nenhum rastro, porque não haverá nenhum meio de exigir que ele efetue um *logoff* (a última página solicitada, antes que o usuário saia da página).

3.3.3 Requisitos de Personalização

Cabreira (2001) destaca alguns requisitos de personalização e técnicas de monitoramento que fornecedores de subsídios ao usuário:

- reconhecimento de revisitas: a fim de personalizar uma sessão da *web*, é essencial ter acesso ao conhecimento anterior sobre o usuário ou o membro da família. Esse conhecimento anterior é em parte obtido a partir de visitas prévias a seu *site da web*;
- interface de usuário e personalização de conteúdo: o primeiro aspecto da personalização é a seleção da interface do usuário. Para tanto, é preciso identificar o tipo e a versão do navegador do usuário a fim de evitar características de conteúdo incompatíveis. Uma vez que se tem conhecimento de visitas anteriores do usuário, é possível personalizar o conteúdo e dar destaque aos itens nos quais ele tem mais interesse;
- vendas colaterais e por impulso: conhecendo o cliente por suas visitas anteriores, é possível lhe sugerir itens correlatos ou outros itens quaisquer;

- filtragem colaborativa ativa: ocorre quando o cliente, uma vez solicitado, evoca uma sugestão de compras futuras, provavelmente em troca de um desconto ou de um presente;
- eventos de calendário e de estilo de vida: alguns eventos de vendas podem ser correlacionados com certas épocas do ano. Os eventos de estilo de vida também são importantes;
- localização: o conhecimento prévio do cliente permite personalizar um site para sua nacionalidade e idioma. O grau de localização pode variar muito, podendo incluir o idioma, formatos de data e hora, moedas e alternativas de remessa.

3.4 Compreendendo a Seqüência de Clique como Fonte de Dados

Para Kimball (2000b), os dados a serem armazenados durante a navegação no *web site* vêm de duas fontes. Da primeira, vêm os dados de *Clickstream*, contidos inerentemente aos protocolos da *web* e armazenadas nos *logs* do *web server*. Da segunda, vêm as informações adicionais sobre as atividades do usuário, a partir do momento em que ele entra e inicia uma sessão na *web site.* Essas atividades são capturadas por aplicações como entrada de pedido, pesquisa no *web site*, visualização de progressão do pedido, entre outras.

A identificação das informações deve ser feita pela sessão e, se for o caso, por algo que identifique o usuário, mesclando e unificando as informações de uma forma coerente no *Data Webhouse*. Apesar de se registrarem informações por usuário, é preciso manter sua privacidade – sua identificação será feita por números gerados e nunca conterá seu endereço eletrônico, número do seu cartão de crédito ou qualquer outra informação particular.

Segundo Kimball (2001), as pessoas não desejam identificar-se ao utilizar a *web* e, quando são forçadas a isso geralmente não dizem a verdade, elas não fornecem dados verdadeiros ou não utilizam o *site.* Por isso, só se deve requisitar informações privadas quando for realmente necessário, como na finalização de uma compra, por exemplo. Essas informações privadas, mesmo se conhecidas, nunca devem ser carregadas no *Data Web*house, mas em um banco de dados à parte, com toda segurança necessária. O motivo dessa separação é o fato de esse tipo de informação não poder ser usado para análise de comportamento e somente para informações necessárias para o sistema transacional.

Farias (2001) destaca a importância de poder construir ferramentas que analisem por completo o *clickstream* como fonte de dados e que formem um *Data Mart*. Para a construção dessas ferramentas, os projetistas do *Data Webhouse* precisam ter amplo conhecimento de diversos aspectos tecnológicos. Farias (2001) destaca algumas experiências deste trabalho:

- plataforma *web* (cliente servidor e múltiplas camadas);
- servidor proxy e cache de browser;
- web server (logs, distribuição de web servers);
- técnicas de sincronização;
- HTML, DHTML, uso de *tags*;
- geração dinâmica de páginas (ASP, JSP, CGI, etc.);
- *cookies* e outros sistemas de identificação.

O conhecimento desses itens possibilita a construção de um *Data Mart* em relação à sua estrutura e a construção de uma ferramenta que processe os dados de c*lickstream* para alimentar o *Data Mart*, seguindo os objetivos e as questões descritas nessa revisão.

3.4.1 *Logs* do Servidor da *Web*

Para Kimball (2000d), todos os servidores da *web* têm a capacidade para registrar interações de clientes em um ou mais arquivos de *log* ou banco de dados ou, ainda para canalizar/direcionar as informações de *log* para outro aplicativo em tempo real.

Dessa forma, o padrão original para *logs* de servidor da *web* é o formato *CLF (Common Log Format)*. Esse padrão é composto por vários elementos de dados (parâmetros registráveis). Os dados de *log* do servidor da *web* são a fonte primária da seqüência de cliques. Toda vez que o servidor da *web* responde a uma solicitação de http, uma entrada é feita no arquivo de *log* do servidor da *web.*

Embora uma entrada seja feita para cada resposta de serviço, o servidor poderá estar mantendo centenas ou até milhares de sessões de usuário, simultaneamente. Assim, os registros individuais que abrangem os rastros da sessão estão dispersos por todo o *log* e devem ser reunidos antes que uma análise completa de sessão possa ser concluída. A seguir,

apresentam-se, no Quadro 3, os elementos de dados de *log* do servidor *web*, com base nas teorias de (NOGUEIRA, 2001):

Quadro 3 – Formato e descrição dos logs do servidor web

VARIÁVEL	DESCRIÇÃO
Host	é qualquer computador na Internet com um nome de domínio com um endereço IP, ou seja, o host é o endereço IP do navegador fazendo solicitação de http. A maioria dos servidores da w*eb* tem a capacidade de transformar esse endereço em um nome de domínio. A conversão de endereço IP pode ser feita fora do servidor da w*eb*, no pós-processamento de seqüência de cliques;
Ident	é a identidade fornecida pelo aplicativo cliente que suporta um protocolo chamado identd (identification daemon);
Authuser	é um ID de usuário passado em uma solicitação feita através do SSP do http;
Time	é a data/hora em que a solicitação alcançou o servidor da w*eb*;
Request	é a primeira linha da solicitação do navegador, normalmente, entre aspas;
Status	é o código de status de três dígitos que o servidor retornar para o navegador;
Bytes	é a contagem de bytes retornada ao cliente pelo servidor;
Referrer	é a URL do servidor de referência;
User-Agent	é o nome e versão do software do cliente/navegador fazendo a solicitação. Essas informações são utilizadas pelo servidor da w*eb* para determinar o conjunto de recursos suportados pelo navegador do cliente e assegurar que a resposta contenha somente itens que possam ser adequadamente interpretados e exibidos pelo navegador;
Filename	é o nome do arquivo é a parte de um URL que especifica o caminho e o nome de um arquivo sendo acessado;
Time-to-Serve	é o tempo para atender à solicitação de http (em segundos); é o IP address – endereço IP de um site da w*eb*;
IP Address	endereço IP de um site da w*eb*;
Server Port	é número da porta TCP/IP em um host, que serviu à atividade de registro (e.g.; porta = 80).
Process Id	é o número do processo filho de servidor da w*eb* que serviu a solicitação;
URL	é o endereço de texto de um objeto específico na w*eb*. Normalmente, consiste em três partes: um prefixo que descreve o protocolo TCP a utilizar para recuperá-lo, um nome de domínio e o nome de um documento.

3.4.2 Cookies

Para Almeida (2001*), cookie* é uma informação que um servidor *web* pode armazenar, temporariamente, junto a um *browser.* Um exemplo de uso desses elementos no comércio via Internet, seria o de alguém entrando em uma loja virtual de CDs, fazendo várias seleções para compra e, em seguida, indo navegar em outros *sites*. Ao voltar ao *site* de venda de CDs, todas as suas seleções teriam sido mantidas e ele poderia então fechar sua compra ou fazer mais aquisições.

As informações são guardadas pelo *browser* e não pelo servidor *web,* o que não deixa de fazer sentido. Ficaria muito mais difícil para um servidor se lembrar dos milhares de *browsers* que o acessaram recentemente e exatamente o que cada um deles fez ou selecionou.

Os *cookies* são enviados para o seu *browser* e mantidos na memória. Ao encerrar a sua sessão com seu browser, todos os cookies que ainda não expiraram são gravados em um arquivo (*cookie file*).

Muitas pessoas julgam que os *cookies* possam ser usados pelo servidor para obter informações a seu respeito ou invadir o seu disco rígido para obter dados a partir de lá. Esse pensamento não procede, pois todas as informações gravadas em um *cookie* são informações que alguém forneceu, voluntariamente, ao servidor, de uma forma ou de outra.

Para criar um *cookie,* o servidor *web* envia uma linha de cabeçalho HTTP em resposta a um pedido de acesso a uma URL solicitada pelo *browser: Set-Cookie: NAME=VALUE; expires=DATE; path=PATH; domain=DOMAIN_NAME; SECURE NAME* é o nome do valor que se está armazenando *no browser, VALUE* é o dado real sendo armazenado no *cookie. DATE* é a data na qual este *cookie* irá expirar (KIMBALL, 2000f).

DOMAIN indica um computador ou rede na qual esse *cookie* é válido. Computadores fora desse domínio não conseguirão ver esse *cookie.* A diretiva "*secure"* indica que o *cookie* somente será transferido sobre conexões seguras (https) e nunca sobre uma conexão *http* normal. De todos esses campos, apenas o campo *NAME* é obrigatório.

Dessa forma, sempre que um navegador solicita uma *URL* a um servidor que nele tenha criado *cookies,* anteriormente, é incluída, juntamente com a *URL,* uma linha listando todos os *cookies* existentes. Essa informação será então utilizada pelo servidor *web* para dar continuidade a transações iniciadas, anteriormente. Essa linha possui um formato do tipo: *Cookie: NAME=VALUE; NAME=VALUE;* (KIMBALL, 2000f).

3.5 Data Mart de Clickstream

Segundo Kimball (2000b), Data Mart de *clickstream* é uma tecnologia que se constitui em modelar tabelas de fatos, ou seja, os tipos de dados desejados de seqüências de cliques capturadas do servidor *web*. É também uma conexão entre *Data Mart* e *Data Webhouse* corporativo. Essa atividade de modelagem consiste, porém em pensar e levantar possíveis dimensões que são relevantes para se caracterizar com os dados do *clickstream*. Com base nestas dimensões e fatos forma-se um esquema to tipo de uma estrela. Então o conjunto desse esquema estrela da se nome de *Data Mart Clickstream*. Com as técnicas apropriadas, porém, o esquema estrela é ligado por meio do uso de dimensões comuns. A tabela de fato mais importante no *Data Mart* de seqüência de clique normalmente é a tabela de fato por sessão. A sua importância é devido ao auxilio que ela fornece em diversas análises sem comprometer a performance e exigir espaço de armazenamento em demasia.

3.5.1 Modelo Dimensional Clickstream

Kimball (2000b) apresenta um modelo de esquema estrela ilustrando um *Data Mart de Clickstream*. As dimensões descritas são Data Universal, Hora Universal, Data Local, Hora Local, Usuário, Página, Evento e Sessão.

As tabelas de datas foram separadas das tabelas de horas, como em outros casos similares, pois têm sentidos distintos. A dimensão data tem o significado de um dia dentro de um calendário, um mês, um dia de semana, uma estação, representados por atributos textuais. A hora é simplesmente a representação de um instante dentro de determinado dia, sem atributos, a não ser que, em um caso específico, se deseje seccionar o tempo em intervalos determinados. Além disso, é improvável que ocorram fatos para todas as combinações possíveis de datas e horas, justificando a representação de uma dimensão com data-hora.

A primeira atividade para modelar um *Data Mart Clickstream* é definir os atributos das possíveis dimensões que são relevantes para se caracterizarem dados capturados. Com base nessas dimensões, forma-se o esquema estrela (Figura 6), apresentado por Kimball (2000b). Cada esquema utiliza parte das dimensões disponíveis, de acordo com a informação desejada do modelo.

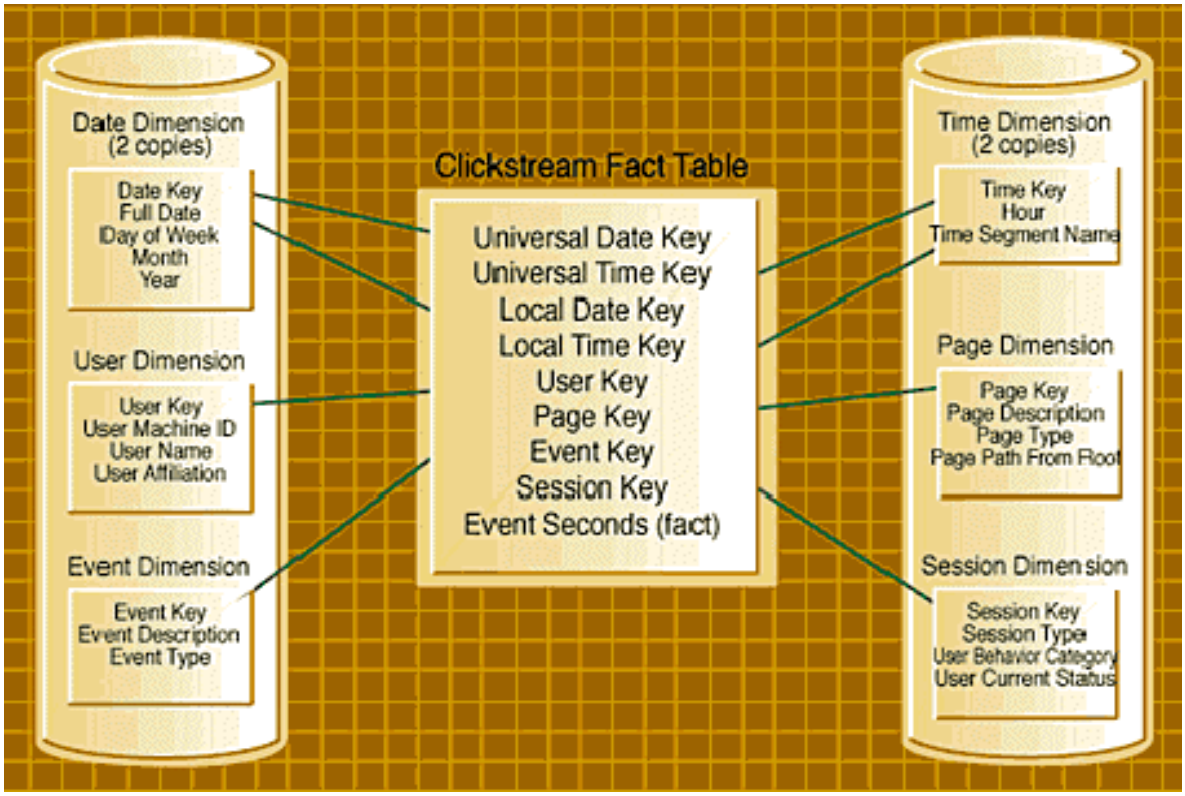


FIGURA 6 - MODELO DIMENSIONAL DATA MART 'CLICKSTREAM' - FONTE KIMBALL (2000B)

Neste contexto, o modelo dimensional acima, deve levar em consideração alguns aspectos apresentados abaixo:

- aspectos relevantes do negócio;
- origem dos fatos a serem medidos;
- granularidade das tabelas de fatos;
- dimensões necessárias para cada tabela de fatos;
- um plano para construir essas dimensões por intermédio da empresa;
- todos os fatos numéricos a serem incluídos nas tabelas de fatos;
- um plano para construir as tabelas de fatos por meio da empresa.

As seções, a seguir, mostram nove dimensões, além de duas tabelas de fatos e uma tabela de fato com as agregações possíveis. Além desses dados, os projetistas podem incluir novas tabelas de acordo com as suas necessidades, porém as tabelas apresentadas normalmente são básicas e farão parte do esquema estrala.

3.5.2 Dimensão Tempo

O *Data Mart de clickstream* também possui a dimensão tempo. A dimensão tempo, na sua maior granularidade normalmente tem um registro para cada dia do calendário, ou seja, é a dimensão data. Se for necessário armazenar a hora do fato, cria-se a dimensão hora, isto é, o tempo é dividido em duas dimensões.

Na dimensão data, é possível registrar atributos como feriados, estações, dias de trabalho, períodos fiscais e outros específicos do negócio. Os outros atributos dessa dimensão são aquelas já tradicionalmente colocadas, como o mês, semana e dia nos diversos formatos, entre outros. A dimensão data do *Data Webhouse* segue os padrões de qualquer *Data Warehouse*. Normalmente, é uma dimensão pequena e a sua chave é uma chave substituta (surrogate key).

Já a dimensão hora pode ser diferente. Num *Data Warehouse* a granularidade da dimensão hora é decidida pelos critérios da realidade. Num *Data Webhouse*, não há sentido, por exemplo, a granularidade da dimensão tempo ser maior que segundos, como milissegundos, por exemplo. Isso se deve ao fato de que a *web* tem um problema intrínseco de sincronização, pela sua distribuição geográfica e pela escalabilidade[4].

Dessa maneira, a informação não é consistente e não se torna útil. Além disso, se for necessária alguma ordenação, pode-se usar outro campo da tabela de fatos, como um campo de seqüência. Então, normalmente se escolhe granularidade em segundos. A dimensão hora também é uma dimensão pequena e possui uma chave substituta.

Uma última observação a ser feita é sobre o fato de uma dimensão tempo, seja data ou hora, não apresentar qualquer indicação de fuso horário, e este é indispensável pelo fato de o alcance da *web* ser global; porém, essa informação irá aparecer de outra forma na tabela de fatos.

[4] Escalabilidade, entende-se por sistemas potencialmente capazes de um aumento gradual do seu poder de atender à demanda, sem prejuízo significativo de prejuízo (Krüger (2001).

3.5.3 Dimensão Cliente

A dimensão cliente também é uma dimensão convencional de *Data Warehouse*, porém os dados dela no *Data Webhouse* não são tão fáceis de serem conseguidos, devido às questões de privacidade e anonimato na utilização da *web*. Esse é o maior desafio na construção de um projeto de *Data Webhouse,* pois se essas informações são conseguidas com qualidade, tem-se a análise muito facilitada.

No caso do *Wehouse,* a dimensão cliente pode ser na verdade a dimensão usuário, visitante ou a dimensão máquina, mas essas informações dependem do nível de detalhe dos dados que serão carregados. Existem grupos de campos de acordo com o nível de detalhe conseguido, o primeiro deles é constituído por campos que sempre são conhecidos por meio da navegação:

- chave do cliente: chave substituta;
- tipo de cliente: descreve o grau de identificação conhecido e desconhecido, regular, não-aplicável, IP fixo, IP variável, cookie fixo, cookie variável, cliente identificado e cliente não identificado;
- endereço do provedor de acesso: multivalorado, o cliente pode-se conectar de casa, do trabalho, etc;
- identificação do Cookie: identificação gerada para o cliente, se possível será a mesma para casa, trabalho, etc;
- data da última alteração e motivo da última alteração: campos de controle que são usados quando existem dimensões que variam lentamente.

O segundo grupo de campos presume que algumas informações de nome e de localização sejam conhecidas e por meio delas se consegue identificar os clientes na navegação, quer dizer, uma chave de cliente pode ser gerada e não é simplesmente uma identificação por *cookie.* Alguns campos são: pseudônimo, nome completo ou parcial, nacionalidade, sexo, cidade, estado e país. O terceiro grupo é o maior nível de detalhe, em que constam: tipo de cliente (residencial ou comercial), etnia, ocupação, empresa, departamento, função, telefone, fax, e-mail, idade, renda, estado civil, interesses, língua (para personalizar mensagens).

Na dimensão cliente, existe uma questão que deve ser tratada com muito cuidado, inicialmente, os campos dessa dimensão devem ser escolhidos de acordo com a realidade, ou seja, devem ser colocados somente os campos que são necessários e que podem ser adquiridos na navegação junto ao cliente. Nesse caso, essa dimensão pode crescer muito mais rapidamente que num *Data Warehouse* convencional, em que a dimensão cliente já é grande. Em comum, tem-se o fato de que é uma dimensão que varia lentamente (*slowly changing dimension*).

Para tratar dimensões que variam, lentamente, há três abordagens já conhecidas, a abordagem a ser utilizada deve ser bem escolhida: tipo 1: sobrescrição, tipo 2: criação de novo registro ou tipo 3: criação de campos para valor inicial e valor atual.

Se houver alguns campos que variam e outros não, ou se o interesse de acompanhar o histórico for de alguns campos, é possível separá-los numa minidimensão demográfica. Um grupo de campos que pode estar numa minidimensão demográfica são os seguintes:

- data da última compra: para saber se é ou não recente;
- freqüência das compras;
- intensidade: número de compras ou valor total;
- tempo de cliente: faz quanto tempo, projetado, é cliente do negócio;
- grupos: campos usados na clusterização;
- perfil: campos de perfil;
- crédito: campos relativos ao crédito;
- retorno de produtos: taxa ou propensão a devoluções;
- uso de suporte on-line / suporte telefônico: taxa de uso.

3.5.4 Dimensão Página

A dimensão página representa o contexto de páginas *web*. O uso precisa ser flexível suficientemente para acompanhar a evolução da *web*. Um exemplo dessa flexibilidade de páginas estáticas que evoluem para páginas dinâmicas, possibilitando a utilização de frames. Isso não significa que a granularidade será ao nível de cada página gerada dinamicamente,

mas é necessário que o mecanismo seja flexível para facilitar manutenções e evoluções na dimensão.

Normalmente, estabelece a identidade das páginas estáticas por sua própria identificação, enquanto as páginas dinâmicas são agrupadas por tipo ou funcionalidade. Quando a página estática ou a definição da página dinâmica é alterada, faz-se necessária a atualização da dimensão página. Como na dimensão cliente, novamente se enfrenta um processo de decisão na forma de alteração da dimensão (tipos 1 a 3), de acordo com o que se necessita guardar e como se deve guardá-lo.

De qualquer forma, precisa-se definir uma codificação para a identificação. Os campos normalmente existentes nessa dimensão são os seguintes:

- chave da página: chave substituta;
- origem da página: estática, dinâmica, desconhecida, corrompida, não-aplicável;
- função da página: portal, busca, descrição de produto, institucional, etc;
- modelo de página: esparsa, densa, etc;
- tipo de item: código do produto, ISBN de livro, etc;
- tipo de gráfico, animação ou som: GIF, JPG, tamanho pré-definido, etc;
- arquivo da página: nome do arquivo Html, detalhes do CGI, ASP, etc;

3.5.5 Dimensão Evento

A dimensão evento descreve acontecimentos particulares em determinadas páginas em determinados momentos de tempo. Alguns eventos interessantes são: abertura, recarga, seleção de link e entrada de dados. Se utilizarem páginas dinâmicas baseadas em XML o número de eventos que podem ser detectados aumenta, pois XML, aumenta a semântica reconhecida pelo W*eb Server*.

Os campos dessa dimensão são a chave substituta, o tipo de evento e o conteúdo do evento, normalmente definidos em meta-tags XML. Essa dimensão é muito pequena.

3.5.6 Dimensão Sessão

A dimensão sessão indica um ou mais níveis de diagnóstico da sessão do usuário. Um contexto local pode estar selecionando produto, enquanto um contexto geral pode estar comparando. Esse estado indica a progressão da atividade que o cliente estiver fazendo.

Além disso, pode-se caracterizar, momentaneamente, o cliente pelos níveis da sessão, por exemplo, novo cliente, preenchendo dados de identificação, cliente cadastrado, cliente confiável, cliente pensando em desistir, cliente padrão, etc. Como esse comportamento muda rapidamente, não há sentido colocá-lo na dimensão cliente, nem criar uma dimensão, ou mini-dimensão já que ele está relacionado à sessão.

Essa dimensão é muito pequena, mas pode ser bastante útil para se fazerem certos tipos de análise, por exemplo, quantidade de visualização dos detalhes do produto antes da compra, quantidade de desistências de compra após visualização de condições de pagamento, quantidade de encomendas não-finalizadas e o ponto de parada. Normalmente, essa dimensão apresenta os seguintes campos:

- chave da sessão: chave substituta;
- tipo da sessão: classificada, não-classificada, corrompida, não-aplicável, etc;
- contexto local: contexto relativo à pagina atual;
- contexto geral: contexto geral relativo à transação efetuada;
- seqüência de ações: descrição resumida das operações feitas na sessão;
- estado de sucesso: sucesso ou fracasso da sessão;
- estado do cliente: padrão, confiável, etc.

3.5.7 Dimensão Referência

A dimensão referência descreve como o cliente chegou à página atual. Essa informação vem do l*og* do We*b Server*: a URL da página anterior e de alguma possível informação adicional.

Os campos desta dimensão são os seguintes:

- chave de referência: chave substituta;

- tipo de referência: Intranet, Internet, máquina de busca, corrompido ou não-aplicável;
- URL: URL que referencia a página atual;
- *site: site* que referencia a página atual;
- domínio: domínio que referencia a página atual;
- tipo de busca: simples ou avançada;
- especificação: expressão útil que busca por texto simples;
- alvo: onde a busca achou a expressão, se meta-tag, cabeçalho ou título, por exemplo. Essa dimensão pode ser bem grande, depende bastante do tamanho dos campos de texto.

3.5.8 Dimensão Produto ou Serviço

A dimensão produto descreve o produto ou serviço que é o assunto da página ou alvo do evento. Essa dimensão não aparece sempre, uma das situações em que não surge, por exemplo, se o *web site* não é de comércio. Essa dimensão é uma das conhecidas pelos projetistas de *Data Warehouse* e, normalmente, essa é uma dimensão que contém um grande conjunto de atributos descritores e hierarquias, por conseqüência, alguns tipos de negócio têm mais dificuldade de compô-la de forma apropriada.

Existem diversas possibilidades de dimensões de produtos e serviços. Um bom exemplo de dimensão de produtos apresenta campos como chave substituta, tipo de produto, fabricante, marca, categoria, código de barra, departamento, tipo de sistema, empacotamento, dimensões físicas, custo, preço, entre outros. Dessa forma, a dimensão normalmente é grande. Um bom exemplo de dimensão de serviços apresenta campos como chave substituta, tipo do serviço, código do serviço, descrição, categoria e setor. Dessa forma, a dimensão é bem pequena.

3.5.9 Dimensão Causal

A dimensão causal descreve as condições do mercado na época em que ocorre o fato, tentando indicar pistas ou fatores causais que podem explicar o interesse do cliente na empresa ou nos seus produtos e serviços. Na seção anterior, citaram-se dois tipos de decisão que tinham relação com fatores causais, detecção quando anúncios, e detecção quando agradecimentos e ofertas conjuntas aos clientes, quando estes faziam o efeito desejado.

Esse tipo de dimensão pode auxiliar na análise de retorno de investimentos em marketing de uma forma geral ou mais especificamente, na correlação entre os acontecimentos de fora da empresa com melhoria ou não no seu desempenho. Normalmente, essa dimensão apresenta os seguintes campos:

- chave de causa: chave substituta;
- tipo de causa: específica, nenhuma, corrompida, não-aplicável;
- tratamento dos preços: regular, redução de percentual, redução de valor fixo, 2 por 1;
- tipo de anúncio no jornal: tamanho e freqüência;
- tipo de anúncio na *web*: news, banner, portal, freqüência de exposição, etc;
- tipo de anúncio no rádio: tempo e freqüência;
- tipo de anúncio na TV: tempo e freqüência;
- localização na loja: está-se localizado de maneira mais ou menos privilegiada;
- tipo de promoção: cupom de desconto, cupom de sorteio, produto extragratuito, etc;

3.5.10 Dimensão Entidade de Negócio

A dimensão de entidade de negócio descreve entidades associadas aos fatos do negócio. Essa é uma dimensão que não é pré-definida em relação ao seu conteúdo. São diversas as entidades que podem desempenhar diferentes papéis no negócio, as quais podem ser fornecedores, parceiros, referência, clientes ou qualquer outro papel desejado. Mas essa dimensão se aproveita de todos esses papéis terem o mesmo formato básico de descrição para

então ser ligada aos esquemas estrela da forma que for conveniente. Essa dimensão apresenta campos como chave substituta, tipo de entidade, nome da entidade, categoria da indústria, nome da pessoa de contato primário, telefone, fax, e-mail, *web site*, moeda usada pela entidade e todos os campos de localização física.

Essa é uma dimensão de tamanho variável que pode ser grande dependendo da quantidade de entidades.

3.5.10 Dimensionando o *Data Webhouse*

Dimensionar um *Data Webhouse* significa definir os requisitos de desempenho para sistemas que administram um volume enorme de transações que um *site* da *web* necessita para transformá-lo em um *Data Webhouse*. Diferentemente do servidor da *web*, com necessidade extraordinária de desempenho de pico, o *Webhouse* sofre demandas extraordinárias de capacidade. Trabalha a uma velocidade constante durante horas seguidas, capturando e digerindo dados para as consultas complexas que passam ser solicitada.

Embora o servidor da *web* necessite de uma escala para administrar cargas de pico, a capacidade de computação do *Webhouse* é determinada, principalmente, pela necessidade de servir a carga de ETL, gerada pelo armazenamento em "buffer" da seqüência de clique do servidor da *web* da empresa e pela carga de consulta imposta ao *Webhouse* por atividades analíticas. A carga de rede do *Webhouse* será muito pequena comparada aos servidores da *web*, mas a necessidade de armazenamento em disco será enorme, facilmente na casa dos terabytes, se uma granularidade boa for dimensionada.

Esse modelo dimensional pode gerar muitas informações valiosas ao tomador de decisão, pois elas permitem conhecer melhor seus clientes, informando: quais são os visitantes mais freqüentes, quais as partes mais visitadas, qual a relação entre os melhores clientes, páginas e usuários. Quanto mais popularizadas estiverem as dimensões sessão e usuário, melhores serão os resultados das análises da seqüência de clique do site (KIMBALL, 2000).

3.6 Estudo Metodológico de *Data Webhousing*

De acordo com Barbosa *et al*., (2002) e Kimball (2000f), uma metodologia de *Data Webhouse* deve prever os seguintes passos:

- definir o objetivo e o público alvo;
- analisar os dados não estruturados e a interação do usuário com o *site*;
- elaborar o modelo dimensional dos dados;
- definir área de estagiamento *(Data Staging Área*);
- implementar os processos de extração, transformação e carga;
- escolha de uma ferramenta de *front end* (para mostrar os resultados);
- analise das informações.

Para Batini (1996), apesar dos passos serem apresentados em um formato que sugere uma seqüência de execução lógica, nem sempre será possível concluir, totalmente, uma fase antes de iniciar a seguinte. Por exemplo, na elaboração do modelo de dados é preciso identificar quais são os dados possíveis de serem capturados pelo processo de captura. Assim, uma avaliação prévia do processo de captura é feita ainda durante o processo de modelagem.

3.6.1 Geração dos Dados Operacionais

Segundo Barbosa *et al*., (2002), a geração dos dados da seqüência de clique é efetuada por meio de arquivos de *logs* gerado automaticamente pelo servidor, ou ainda pela confecção de rotinas próprias para captura de dados no servidor que recebe, de cada página do *site*, a fim de complementar as informações que possivelmente os tomadores de decisão necessitarão.

Todavia, dependendo da estrutura do *site,* é possível utilizar esse serviço de captura de dados de seqüência de clique juntamente com o arquivo de *logs.* De qualquer forma, basta, para isso, que em cada uma das páginas do *site* seja incluído o código que realiza a coleta de dados e que os envia ao serviço de captura. A forma de coleta de dados mais tradicional, realizada pela maioria dos sistemas de análise de *log*, ocorre por meio da leitura e interpretação do arquivo de registros de *log* do servidor *web,* apresentando como resultado uma análise que levam em consideração as páginas requisitadas ao servidor.

3.6.2 Arquitetura de uma Solução

Farias (2002) destaca que por meio da Figura 7 uma forma de desenvolver um *Data Webhouse* simples e os principais passos para análise de seqüências de clique. Na implementação, são tratados todos os processos, exceto a obtenção de dados de outras fontes.

A Figura 7 apresenta uma visão dos passos a serem gerados:

1) geração de dados operacionais;

2) preparação dos dados na *Data Staging Área*;

3) publicação dos dados no *Data Webhouse* e a interação entre esses passos com o ambiente externo, objetivando a análise das consultas com ferramentas de *OLAP*.

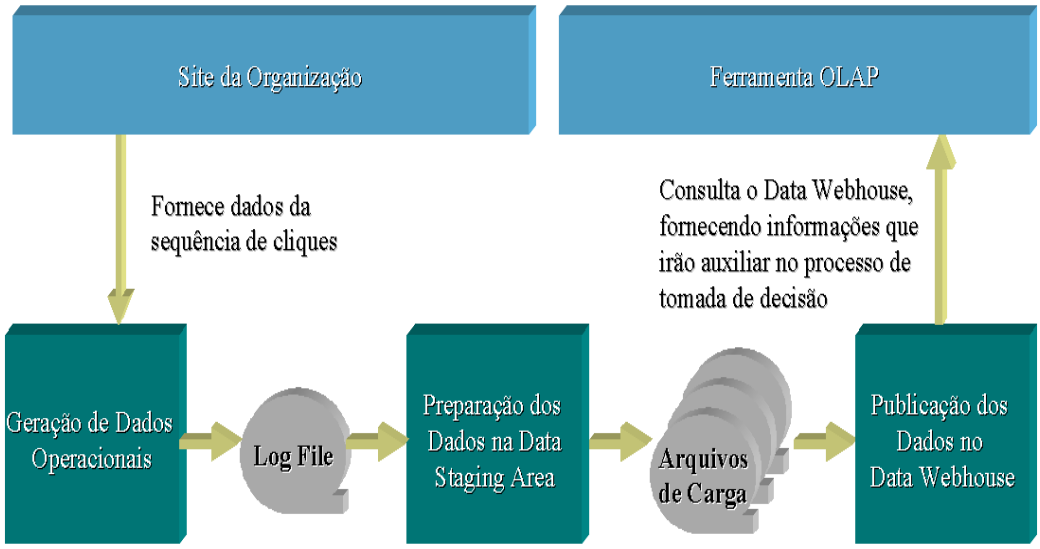


FIGURA 7 – ARQUITETURA DE UM DATA WEBHOUSE SIMPLES – FONTE ADAPTADO DE FARIAS (2002)

Segundo Farias (2002), a interface entre os passos é realizada por meio de arquivos texto de formato padronizado. Os passos 1 e 2 se integram por meio de um arquivo de *log* em formato proprietário. Os passos 2 e 3 geram um conjunto de arquivos texto (arquivos de carga) a serem utilizados pelo passo 3, na atualização do *Data Webhouse.*

Essa solução de interface visa a trazer uma independência entre as fases, permitindo que cada um possa evoluir, separadamente, sem que isso cause um impacto direto nas demais etapas do projeto.

3.6.2.1 Área de Estagiamento

Segundo Kimball (2000f), *Data Staging Área* é o local onde as informações são extraídas dos sistemas transacionais para armazenar informações temporárias. Nessa fase, também é processada a etapa de transformação e armazenada para carga no *Data Webhouse*.

A extração é um processo de seleção das informações brutas, pois a mesma trata de uma organização dos dados no modelo que originará *o Data Webhouse*, porém essa extração acessa os sistemas transacionais; no caso do *Webhouse,* acessa arquivos logs do servidor de Internet. Os dados capturados do servidor de Internet serão utilizados para serem incluídos em um modelo de dados intermediários que serão manipulados pelo processo de transformação.

3.6.2.2 Publicação das Informações no Data Webhouse

De acordo com Farias (2002), esse processo visa a carregar os dados transformados da área de estagiamento para um banco de dados *Data Webhouse*, inserindo dados de histórico aos novos usuários e acrescentando-os também aos usuários existentes. A Figura 8 apresenta uma visão geral do processo de carga, em que são lidos os arquivos gerados pelas fases de extração, transformação e publicação.

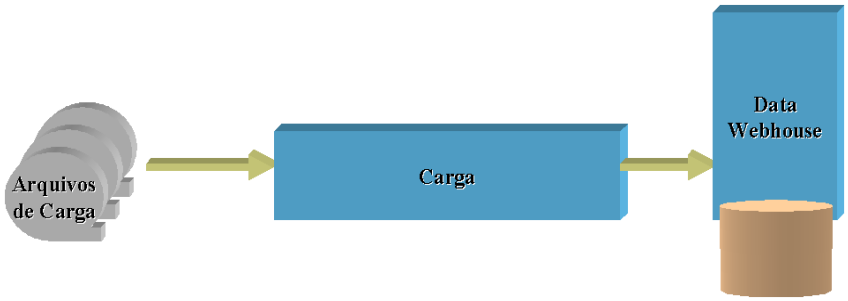


FIGURA 8 - VISÃO GERAL DO PROCESSO DE CARGA - FONTE – ADAPTADO DE FARIAS (2002)

Os processos de cargas deverão ser efetuados por meio da utilização de rotinas desenvolvidas pelos programadores do projeto, esses programas deverão ler os arquivos de logs e gravar em tabelas temporárias. Para fazer a carga dos arquivos gerados na fase de extração e transformação de dados, é preciso que as tabelas estejam geradas no banco de dados o qual, possivelmente, será publicado no *Data Web*house.

3.6.3 Uma Visão Geral dos Processos de Extração, Transformação e Carga

Segundo Cielo (2001), a extração, transformação e carga (ETL) é a etapa mais crítica do *Data Webhouse,* pois envolve a movimentação consistente dos dados da origem (sistemas transacionais) para o destino (servidor de publicação dos dados).

Dessa forma, as literaturas apresentadas na forma de revisão bibliográfica, apresentam o caminho necessário para a extração de informação nos arquivos de *logs* do servidor de Internet. O primeiro passo realizado foi extrair os dados dos arquivos de *logs*, porém esse processo de extração na base capturada implica a identificação das fontes de informação (sistemas transacionais) que deverão ser consultadas para obtenção dos dados que mais tarde, serão feitas as cargas nas dimensões e nas tabelas de fatos.

No processo de extração, alguns pontos devem ser analisados:

- o processo de extração de cada elemento (atributo ou fato) deverá seguir regras bem definidas nos metadados;
- os dados necessários para compor determinado elemento poderão estar armazenados em diferentes fontes de dados, em diversos formatos, implicando mudança de tecnologia;
- é necessário prever a existência de valores padrão a serem preenchidos, quando o valor de algum elemento não puder ser definido.

Dessa forma, Kimball (2000f) destaca que, a princípio, os dados extraídos deverão ser armazenados em uma área de armazenamento temporária (*Data Staging Area*), para serem manipulados pelos processos subseqüentes. Essa área de armazenamento de dados poderá ter um repositório de metadados que irá armazenar as informações sobre as regras de mapeamento dos dados. Caso esse repositório não seja implementado, essas informações deverão estar presentes no repositório de metadados do servidor de publicação de dados do *Data Warehouse.*

Kimball (1998) ressalta, ainda, que as informações armazenadas na área de armazenamento temporário só poderão ser manipuladas pelos processos de extração, transformação e carga, não podendo, de forma alguma, haver consultas diretas a esses dados para outras finalidades.

Já Farias (2002) também define que, uma vez extraídos os fatos e atributos, é preciso transformá-los e esse processo consiste na realização de várias atividades:

- a solução mais adequada é a utilização de chaves substitutas, ou seja, chaves geradas especificamente para diferenciação dos registros das dimensões no *Data Webhouse,* isolando a estrutura do *Data Warehouse* da influência das alterações que possam ocorrer no ambiente de produção. A proposta de Kimball e Merz (2000f) é a utilização de limpeza/formatação dos dados: eliminação das inconsistências e informações que não atendam às regras de controle de qualidade estabelecidas para aceitação dos dados no *Data Warehouse* e da conversão dos dados no formato desejado;

- o tratamento das dimensões, a princípio, as informações armazenadas nas dimensões não devem ser alteradas. Sempre que surgirem novos valores, estes deverão ser incluídos. Dimensões que se alteram lentamente são dimensões para as quais alguns atributos de um registro são atualizados, e para o processo de análise, o registro continua o mesmo;

- definição de chaves substitutas: os registros de cada dimensão possuem uma chave primária, a qual é exportada para a tabela de fatos, estabelecendo uma ligação entre elas. O impulso inicial é no sentido do aproveitamento das chaves da produção como chaves primárias das dimensões.

- evitar a inconsistência de informações em função do reaproveitamento de chaves na produção;

- evitar uma futura incompatibilidade entre chaves, o que acontece, por exemplo, quando a chave da produção utiliza uma informação fornecida pelo fabricante do produto, e a organização adquire produtos de outro fabricante com um formato de chave diferente;

- o tratamento com os registros nas tabelas de dimensão que representam informações não-existentes na produção. Isso ocorre, por exemplo, numa dimensão produto, no caso de registro de um fato em que o produto não pode ser detectado.

O próximo módulo irá apresentar uma carga de dados em que serão inseridas informações no servidor de banco de dados *Data Warehouse*. Normalmente, em primeiro

lugar são carregadas as dimensões e depois as tabelas de fatos. O processo de carga, na visão de Kimball e Merz (2000f), envolve a execução de uma carga inicial, na qual serão gerados valores para algumas dimensões, como data e hora, que não são extraídas dos sistemas transacionais, e a execução de cargas periódicas são para que os dados publicados contenham os valores mais atualizados.

De acordo com Cielo (2000), existem ferramentas para automatizar esses processos, e a seleção da ferramenta mais adequada varia de acordo com as necessidades de cada organização.

3.6.4 Arquitetura de Processos

Segundo Kimball (2000), os processos de seqüência de cliques podem ser implementados tanto como um aplicativo orientado à transação quanto em *batch*, ou com uma combinação destes. Independentemente da solução adotada, em linhas gerais, as etapas a serem executadas pela arquitetura de seqüência de clique tem o objetivo de carregar as tabelas de fatos que estão descritas na Figura 9. Cada uma dessas etapas será descrita na seção de plantação.

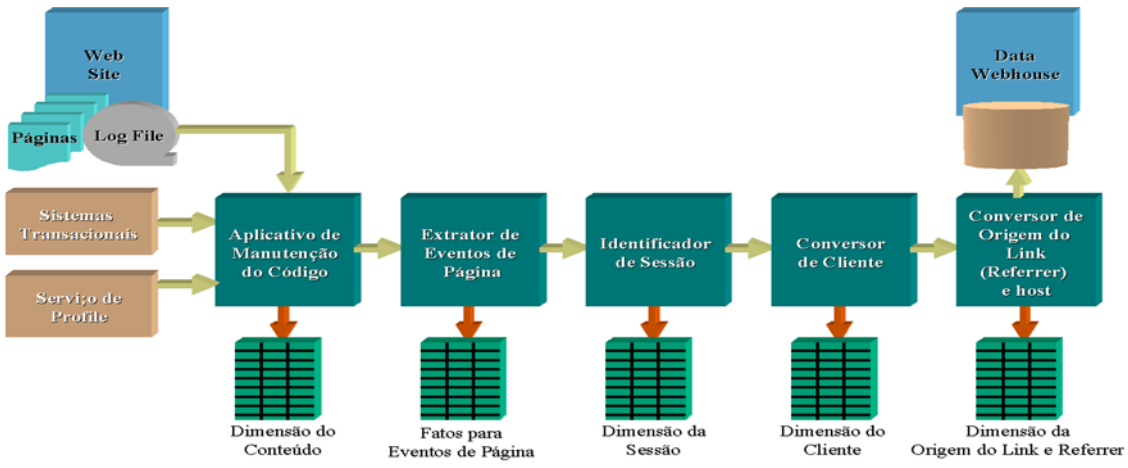


FIGURA 9 - DETALHAMENTO DO PROCESSO DE IMPLEMENTAÇÃO ETL - FONTE FARIAS (2002)

Segundo Barbosa *et al*., (2002), após a análise das informações capturadas do servidor, deve-se passá-las por um processo de transformação e limpeza dos dados, para serem carregados no modelo dimensional e posteriormente publicados no *Data Webhouse*. O processo de publicação compreende três etapas: extração, transformação e carga. Esse

processo representa uma ponte entre os dados brutos capturados e as informações publicadas no *Data Web*house. Esse passo costuma ser o mais difícil de ser efetuado, pois maior é a gama de alternativas disponíveis para implementação e, conseqüentemente, maiores são as possibilidades de investimento.

Para Farias (2002), o processo de *ETL* tem como objetivo operar como um pós-processador de seqüência de clique, preparando os dados capturados para serem carregados no banco de dados. A implementação do processo de *ETL* pode ser subdividida nas seguintes atividades principais:

- determinar quais são as fontes de dados;
- definir os passos necessários para transformar os dados;
- preparar os dados num formato compatível com as tabelas geradas do modelo proposto;
- carregar os dados.

A Figura 10, mostra os processos de extração e transformação para a carga dos dados para o modelo dimensional definido.

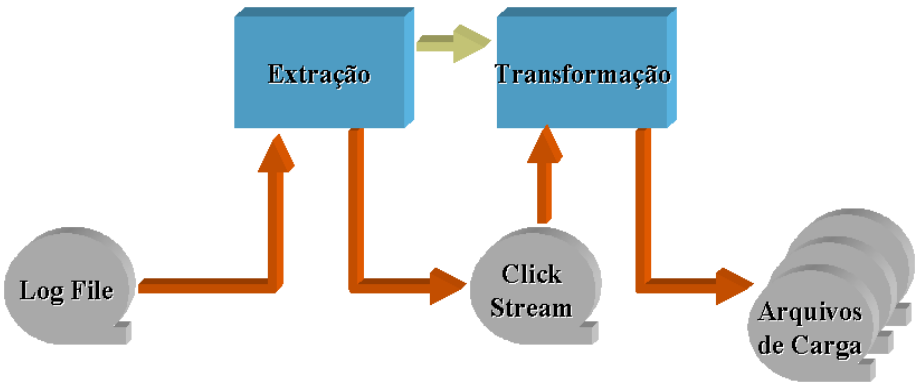


FIGURA 10 - PROCESSO DE EXTRAÇÃO E TRANSFORMAÇÃO – ADAPTADO DE KIMBALL (2000)

Para Kimball e Merz (2000f), a fase de transformação dos dados é extremamente técnico, uma vez que envolve rotinas de programação e analise de requisitos que o sistema irá disponibilizar através do modelo dimensional definido. Antes de montar os arquivos para carga, são estabelecidas as chaves para cada uma das dimensões do *Data Webhouse*. Segue, como exemplo, o Quadro abaixo com as chaves de dimensão e responsabilidade de atribuição para cada tabela do modelo proposto.

Quadro 4 - Dimensões do quadro de evento de página - fonte: (KIMBALL 2000)

CHAVES DE DIMENSÃO	RESOLVIDA POR
Chave de Data	Extrator de Evento de Página
Chave de Cliente	Identificador de Sessão
Chave de Evento	Extrator de Evento de Página
Chave de Referência de Origem	Conversor de Host/Origem do link (referrer)
Chave de Causa	Marketing (inserção manual)
Chave de Data/Hora	[Nativo]
Chave de Página	Conversor de Conteúdo
Chave de Sessão	Identificador de Sessão

3.6.4.1 Implantando o Processo de Carga

De acordo com Kimball (2000), a carga visa carregar as informações das dimensões e tabelas de fatos da área de armazenamento temporário para o servidor de publicação dos dados do *Data Webhouse*, conforme apresentado na Figura 11.

A implementação do processo de carga pode ser subdividida nas seguintes atividades principais:

- determinar as características do processo de atualização dos dados publicados (carga total ou incremental);
- determinar a periodicidade de execução da carga;
- definir os passos necessários para carregar os dados.

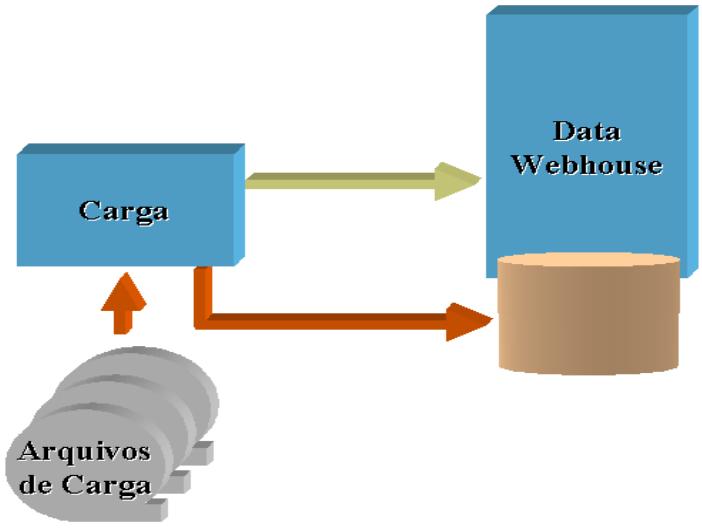


FIGURA 11 - PROCESSO DE CARGA – FONTE BARBOSA *ET AL*. (2002)

Considerando-se que o objetivo do *Data Web*house é análise das interações dos clientes com o site ao longo do tempo, a carga deverá ser incremental. Uma primeira carga deverá ser executada para gerar o conteúdo das dimensões Data e Hora e povoar as demais dimensões que não dependam das informações da seqüência de clique, como a dimensão página, por exemplo.

Se o processo de carga estiver realizando a verificação da integridade referencial das informações que estão sendo carregadas, as dimensões deverão ser carregadas antes dos fatos.

3.6.5 Analisando o Comportamento dos Usuários

Contudo, para Kimball e Merz (2000f), o simples registro de todas as ações dos usuários, numa base de dados, não é suficiente para analisar o seu comportamento. Os dados sem tratamento não mostram os caminhos e as opções realizadas pelo usuário. Entende-se como uma ação, toda a solicitação de recursos do *site,* normalmente, a requisição de páginas ou execução de operações fornecidas como serviço no *site*, tais como operações para compra de produtos.

Uma descrição útil de comportamento do usuário é necessária. Essa descrição não se pode resumir a poucas variáveis ou a descrições gerais. É desejável ter-se certo nível de detalhe. Mesmo com um nível maior de detalhes, é possível que diversas páginas estejam associadas a uma mesma descrição de comportamento.

Saber traduzir qual o comportamento adotado pelo usuário, em função das ações realizadas por ele mesmo, é a solução principal para análise da seqüência de clique. A forma de tradução das ações para comportamentos influenciará diretamente na qualidade do *Data Webhouse*, isto é, quanto melhor for a solução adotada para deduzir o comportamento do usuário em função das ações por ele realizadas, mais correto é o resultado obtido.

3.7 Considerações Finais

Este capítulo teve como objetivo apresentar uma visão geral sobre a tecnologia *Data Webhouse* e apresentar conceitos da literatura, mostrar suas características, viabilidades de uso e sua importância para as tomadas de decisões. Kimball e Merz (2000), apresentam essa tecnologia como uma “poderosa” ferramenta para as organizações, pois ao coletar dados

valiosos sobre os usuários, ela pode auxiliar a criar melhores serviços, no que se refere, especificamente, ao monitoramento de *sites*, bem como das seqüências de clique como uma fonte de dados.

Analisaram-se, também, as metodologias, no aspecto tecnológico, concluiu-se que o *Data Webhouse* é mais resultado de uma técnica apurada de preparação cuidadosa do que inovação tecnológica. Por esse motivo, o planejamento e o conhecimento do negócio assumem uma importância vital na elaboração de uma aplicação metodológica.

Dessa forma, a visão apresentada torna-se preponderante, tanto para o entendimento dos próximos capítulos, quanto para o entendimento do modelo que será implementado no próximo capítulo. Na seqüência, será apresentada uma aplicação das técnicas de *Data Webhousing* para Grupos de P&D e como proposta de melhoria um *site* modelo para os Grupos de Pesquisa nas IES.

4 APLICANDO AS TÉCNICAS DE DATA WEBHOUSING SOBRE O SITE DE UM GRUPO DE PESQUISA E DESENVOLVIMENTO

4.1 Considerações Iniciais

Neste capítulo será apresentado como o processo de *Data Webhousing,* introduzido nos capítulos anteriores, foi aplicado sobre o *site* de um Grupo de P&D. Para tanto, foi concebido um protótipo, cujo objetivo foi o de subsidiar o aprimoramento da estrutura e do conteúdo do *site* do Grupo de P&D analisado e, por extensão, aos demais *sites* de Grupos de P&D.

Inicialmente, é apresentado um levantamento realizado sobre alguns *sites* de Grupos de P&D brasileiros, a fim de identificar padrões sobre a estruturação de conteúdo nos *sites* analisados. Em seguida, é apresentada a versão do *site* do Grupo Stela analisada e uma breve introdução sobre a organização do conteúdo desta versão. Por fim, são apresentadas as fases do processo de *Data Webhousing* que foram implementadas na construção de um protótipo com os dados deste *site.*

O protótipo construído visa contribuir para a melhoria da estrutura e do conteúdo de *sites* de Grupos de P&D a partir da análise dos acessos dos usuários desses *sites* e posterior personalização do conteúdo de acordo com as reais necessidades desses usuários. Os indicativos obtidos através do protótipo e a partir da comparação do conteúdo deste *site* com os demais *sites* analisados, subsidiaram as análises que serão apresentadas no próximo capítulo.

4.2 Sites Web de Grupos de Pesquisa e Desenvolvimento

Esta seção abordará as estruturas de *sites* dos Grupos de P&D, cujo objetivo é ilustrar o perfil da pesquisa científica, no Brasil, e demonstrar a situação desses *web sites* no meio científico, quando comparados a outras organizações de ensino e pesquisa, no País.

A maioria dos Grupos de P&D no Brasil possui um *site* na Internet e deseja apresentar suas pesquisas, divulgá-las, expor seus projetos realizados, ou em outras palavras, fazer-se

visível. No entanto, verifica-se que poucas informações essenciais são disponibilizadas e organizadas em seus *sites*. Diante dessa observação, os *sites* de Grupos de P&D, deverão seguir os seguintes objetivos:

- organizar e disponibilizar de forma sistêmica as informações sobre a pesquisa institucional e a produção científica dos pesquisadores do Grupo;
- articular meios junto às agências de fomento, a obtenção de recursos financeiros para viabilizar o desenvolvimento da iniciação científica;
- assimilação do conhecimento existente;
- resultados esperados e seus usuários.

Por outro lado, Nielsen (2000) ressalta que as idéias de competição em busca da qualidade de informação de serviços foram maximizadas em relação à divulgação cientifica na Internet. Além disso, houve apoio estrutural ao *site,* planejando, com base em métodos eficientes, uma navegabilidade personalizada para cada *site*, perfazendo assim os meios mais modernos para a divulgação da Ciência e Tecnologia, no Brasil.

Cabe ressaltar, então, que é viável desenvolver uma metodologia que auxilie os Grupos de pesquisa a disponibilizar sua Produção Científica na *Web,* e contribuir com informações de modo que venha a aumentar a interatividade entre os pesquisadores.

Diante desse cenário, o Quadro 5 demonstrado abaixo apresenta resultados de um levantamento de pesquisa, que teve como objetivo demonstrar uma pré-seleção de resultados estatísticos de maneira como estão estruturados os *sites* dos Grupos de P&D atualmente no Brasil. Essa pesquisa foi uma amostragem aleatória verificado no mês de agosto de 2002, com *sites* de Grupos de P&D de diversas áreas do conhecimento, onde se pode observar como se apresentam estruturados doze *sites* escolhidos aleatoriamente, através de ferramentas de busca na Internet.

Nesta amostragem ilustrada no Quadro 5, apresentam-se os *sites* que foram analisados para subsidiar no levantamento de requisitos da aplicação e propor um novo *site* no capitulo 5 deste trabalho. Contudo, esses dados deverão ser transportados para a aplicação através dos *logs* do Grupo de Pesquisa indicado para o estudo de caso deste trabalho, afim de que possa subsidiar o *Data Webhousing* na análise dos resultados. Por meio dessas técnicas, será possível conhecer melhor os pesquisadores parceiros que estão acessando o *site*. De acordo com esses resultados, poderão ser abertas novas perspectivas para melhoria dos *site web* de

seus respectivos Grupos de Pesquisa e, conseqüentemente, o aprimoramento dos seus objetivos.

Quadro 5 – Como estão estruturados os sites de grupo de P&D, no Brasil

Itens Pesquisados nos Sites de de Grupos P&D	1	2	3	4	5	6	7	8	9	10	11	12
Histórico do grupo?	Sim	Sim	Sim	Sim	Não	Sim	Não	Sim	Não	Sim	Sim	Sim
Missão e o perfil do grupo?	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não	Sim
Apresenta equipe?	Sim	Sim	Não	Sim	Sim	Sim	Sim	Sim	Não	Não	Sim	Sim
Apresenta P&D do grupo?	Sim	Sim	Sim	Sim	Sim	Não	Sim	Sim	Não	Não	Sim	Não
As linhas de pesquisas?	Sim	Sim	Sim	Sim	Sim	Sim	Não	Sim	Sim	Não	Sim	Sim
Projetos realizados?	Não	Sim	Sim	Não	Não	Não	Não	Sim	Não	Não	Sim	Não
Titulação do grupo?	Não	Sim	Não	Sim	Sim	Não	Sim	Não	Não	Não	Não	Não
Tese & dissertações orientadas?	Não	Não	Sim	Sim	Não	Não	Não	Sim	Sim	Não	Não	Sim
Recursos utilizados pelo grupo?	Não	Sim	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não
Usuários do grupo?	Não	Sim	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não
Apresenta seu Portfólio?	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não	Não
Disponibiliza o mapa do site?	Não	Não	Não	Não	Sim	Não	Não	Não	Não	Não	Não	Não

Diante dessa verificação, observou-se que os principais indicativos apresentados para serem pesquisados não estão disponibilizados, adequadamente, nos *sites*. Por outro lado, item como linhas de pesquisa, por exemplo, apresenta em 88% dos *sites* com essa informação.

Com os recursos que a *web* possui, o conhecimento poderá ser compartilhado entre os mais diversos Grupos de Pesquisa, no Brasil, e com as técnicas de *Data Webhousing,* é possível medir a audiência do *site* e traçar objetivos que venham ao encontro das necessidades dos Grupos e seus pesquisadores. Uma melhor definição dos itens apresentados no Quadro 5, deverá ser resgatado no final do capítulo 5, quando será apresentado uma proposta de *sites* para os diversos Grupos de P&D brasileiro.

4.3 Site de Grupo de Pesquisa e Desenvolvimento

O Grupo Stela é um laboratório de Pesquisa e desenvolvimento de sistemas de informação e de inteligência aplicada da Universidade Federal de Santa Catarina (UFSC). Formado em 1995, no Programa de Pós-Graduação em Engenharia de Produção, combina a pesquisa acadêmica e o desenvolvimento de tecnologia de ponta com o auxílio de uma equipe de doutores, doutorandos, mestrandos e graduandos que trabalham na criação de novas

metodologias e tecnologias nas áreas de sistemas de informação, inteligência aplicada, em engenharia e gestão do conhecimento (GRUPO STELA, 2001).

Um projeto de relevância nacional resultou numa nova parceria com o CNPq, que consistiu na concepção e no desenvolvimento da Plataforma de Sistemas de Informação para integração dos aplicativos brasileiros, utilizados por pesquisadores, estudantes e gestores de C&T para cadastro de informações, o que originou a construção da Plataforma Lattes.

A Plataforma Lattes compõe-se de um conjunto de sistemas que promove suporte à captação e manutenção dos dados curriculares dos pesquisadores no País, dividindo-se em vários sistemas responsáveis, desde o preenchimento dos dados curriculares pelo pesquisador, através do Currículo Lattes, passando pelos sistemas de recepção dos dados e os sistemas de controle dentro da agência (CNPq) (GRUPO STELA, 2001).

Dessa forma, em 2002, o Grupo avançou em vários projetos, com vistas a ampliar, consolidar e internacionalizar a Plataforma Lattes. Diversos produtos de informação que visam à gestão, o intercâmbio e a divulgação de informações em C&T foram desenvolvidos, são exemplos, os Portais Lattes de C&T, Egressos e Demografia Institucional da Pesquisa. Também foram concluídos os instrumentos de avaliação, divulgação e análise da pesquisa brasileira, mapeada na unidade de Grupo de Pesquisa. Nesse contexto, o Grupo Stela direciona sua linha de pesquisa no desenvolvimento, na formação e extensão nas áreas de Tecnologia da Informação e Engenharia do Conhecimento.

As principais ações do Grupo são:

- formação continuada de graduandos, mestrandos e doutorandos;
- capacitação e transferência de tecnologia em projetos de extensão;
- P&D de produtos na área de tecnologia da Informação;
- P&D nas áreas de Inteligência Aplicada, Engenharia de Software, Portais Corporativos, Gestão de Conhecimento e TI (Tecnologia da Informação) na Gestão Universitária;
- projetos e prestação de serviços especializados nas suas áreas de P&D.

Neste aspecto, o objetivo do Grupo é a organização e a disponibilização das informações sobre a pesquisa institucional e produção científica. Cabe salientar ainda que a versão do *site* analisado (Plataforma Stela, início de 2002), apresentava as prestações de serviço ao Programa de Pós-Graduação em Engenharia de Produção (PPGEP), conforme foi

ilustrado na Figura 13 onde o *site* apresentava as principais informações catalogadas nas suas páginas, que são:

- Stela Net
- Stela Inscrição
- Stela Prof
- Stela Estatística
- Tese e Dissertação
- Informações do Grupo Stela

A Figura 12 apresenta o *site* da Plataforma Stela de 2002 e, em seguida, apresenta uma descrição do mapa do *site* com todas as informações disponibilizadas. Essa apresentação se faz necessária, para que o leitor possa acompanhar a metodologia de desenvolvimento, que será descrita na seção 4.4 deste capítulo.

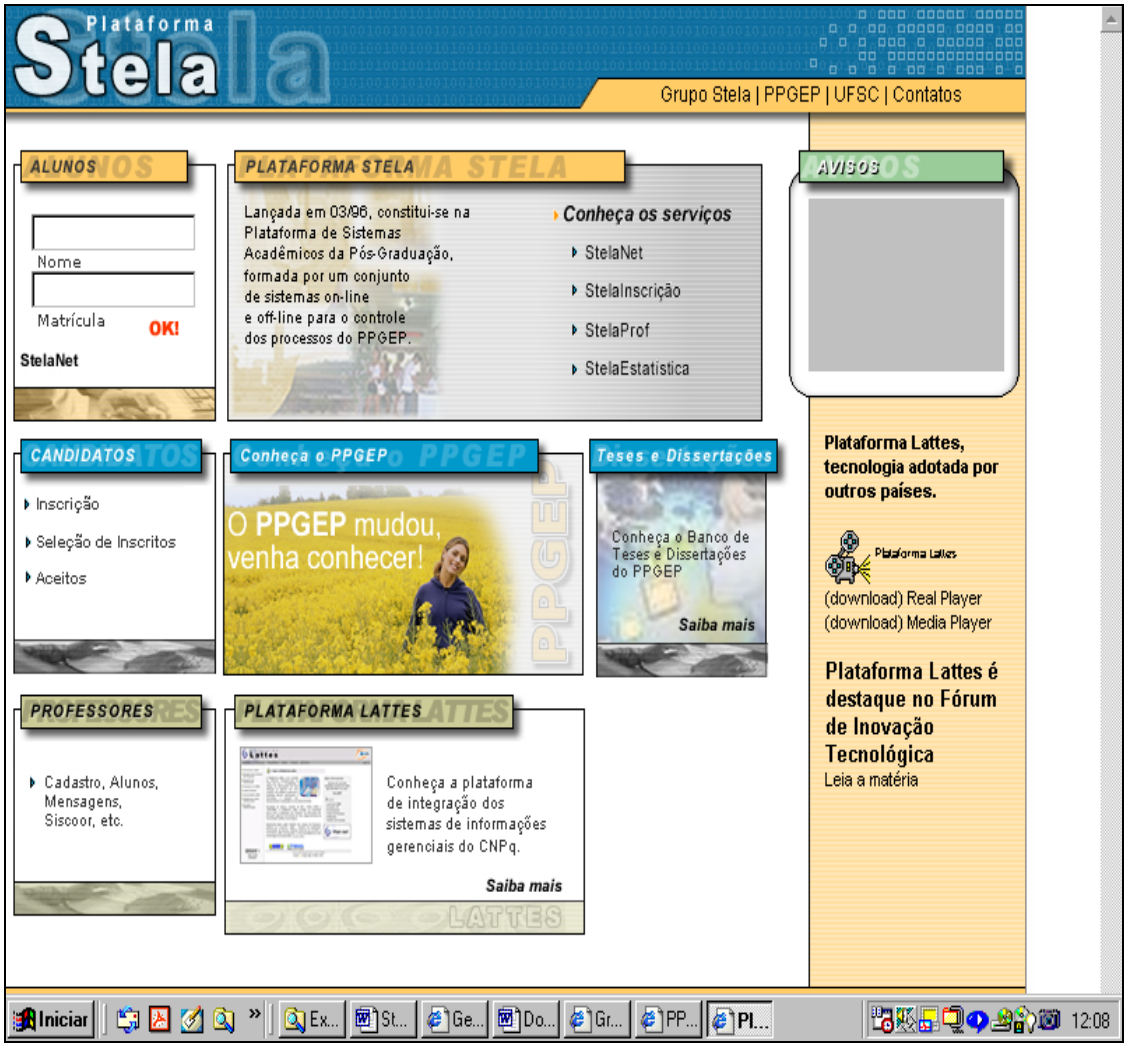


FIGURA 12 – SITE DA PLATAFORMA STELA DE 2002

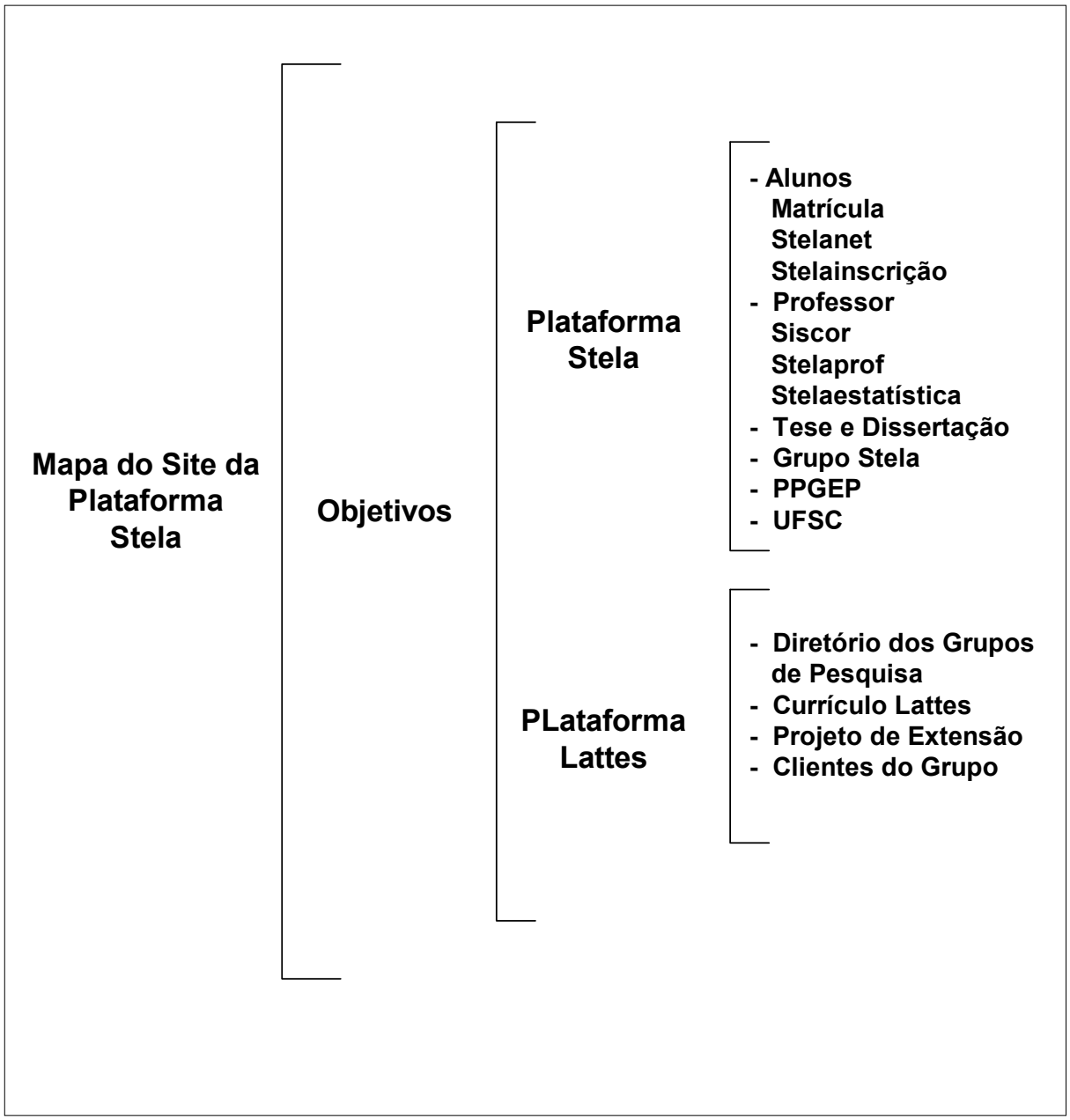


FIGURA 13 – DIAGRAMA DO SITE DA PLATAFORMA STELA DE 2002

4.3.1 Plataforma Stela

A Plataforma Stela compreende todos os serviços, sistemas e projetos desenvolvidos especialmente no Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEP). Dessa forma, conforme a Figura 13 (diagrama do *site*), apresenta duplo objetivo, ou seja, demonstra de forma mais destacada os serviços prestados ao PPGEP e à Plataforma Lattes. Além da Plataforma Lattes, porém mostra outros serviços como: Diretório de Grupos de Pesquisa e Projetos de extensão em geral.

4.3.2 Plataforma Lattes

Com a decisão de ampliação do projeto de integração dos sistemas de informações surgiu a Plataforma Lattes resultado do esforço conjunto de vários órgãos, entre eles, Ministério de Ciência e Tecnologia (MCT), Conselho Nacional de Pesquisa (CNPq), Financiadora de Estudos e Projetos (FINEP) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

A Plataforma Lattes é um dos projetos desenvolvido pelo Grupo Stela, onde além de agregar a integração das bases de dados melhorando o fluxo de informações para pesquisadores, instituições, agências de fomento e órgãos do governo, realiza também trabalhos de extensão na área de sistemas de informações, como o Diretório de Grupos de Pesquisa, no Brasil, projeto do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que visa reunir informações sobre todos os grupos de pesquisa em atividade no país.

O Currículo Lattes é o formulário eletrônico responsável pela coleta das informações que servem de apoio na descrição da pesquisa, no País em nível de indivíduo. Essas informações são, em geral, originadas de pesquisadores ou usuários do CNPq, que requisitam recursos, sejam estes bolsas ou auxílios para projetos de pesquisa.

De um modo geral, a Plataforma Lattes fornece subsídios ao incremento e manutenção da base de dados curriculares do CNPq. Portanto, este projeto tem como uma de suas principais finalidades, efetivar o potencial fornecido através da estruturação da infra-estrutura dos currículos no CNPq, de forma a possibilitar a construção de um repositório comum de currículos a todos os agentes institucionais de pesquisa e desenvolvimento de C&T, no Brasil.

Na Figura 14 abaixo, apresenta-se o *site* atual do Grupo de Pesquisa Stela, totalmente remodelado ainda em 2002, agora nos padrões ideais e estruturado de acordo com os objetivos de Grupos de P&D. Dessa forma, vale registrar que o *site* atual do Grupo foi fruto de elaboração e concepção da equipe de Design e Projetos do Grupo Stela.

Portanto, a pesquisa aplicada, neste trabalho, analisou o *site* da Plataforma Stela 2002, pois, quando da concepção desse trabalho e da captura de dados dos arquivos de *logs*, o novo *site* ainda não estava implantado. Diante desse novo cenário, pode-se dizer que o *site* atual é o mesmo da Plataforma Stela 2002, o qual é o objeto de análise deste trabalho, contudo, a

diferença está em que no *site* da Plataforma Stela 2002 foram separadas as informações do Grupo, em um *site* institucional de P&D e outro *site* com recursos informacionais no PPGEP.

Finalmente, vale destacar que, o *site* atual do Grupo Stela está estruturado, adequadamente, de acordo com os objetivos estudados no capítulo anterior (GRUPO STELA, 2001). A figura 14 representa o *site* atual do Grupo.

FIGURA 14 – SITE ATUAL DO GRUPO DE PESQUISA E DESENVOLVIMENTO STELA

Na próxima seção, serão resgatados alguns aspectos descritos, anteriormente, no capítulo 2 dentro do contexto do protótipo desenvolvido. A proposta para implantação do *Data Mart* seqüências de clique, vislumbra um contexto onde administradores e gestores poderão ter acesso sobre informações relevantes dos cliques dos usuários que navegam em seus *web sites*, vale dizer, o caminho percorrido por eles dentro do *site*, que, certamente, por meio dessas técnicas e ferramentas que serão apresentadas, poderão obter informações importantes para a definição do conteúdo e organização das informações em um novo *site.*

4.4 APLICAÇÃO DE *DATA WEBHOUSING* NA ANÁLISE E REESTRUTURAÇÃO DE SITE DE GRUPOS DE P&D

Esta seção apresentará os processos e as fases de preparação para o desenvolvimento de um projeto de *Data Webhousing*, baseando-se para isso nas metodologias propostas por (BARBOSA, ET AL., 2002), (FARIAS , 2002) e (KIMBALL, 2000f).

Inmon (2001) destaca que a seqüência de clique tem como objetivo suprir as deficiências das fontes de dados tradicionais no ambiente *web*. Com isso as seqüências de clique não são somente mais uma fonte de dados que foi extraída, limpa e organizada no *Data Webhouse*. Elas são, na verdade, uma coleção de fontes de dados, já que existem diversas formas de registrar o comportamento dos usuários e, de acordo com suas necessidades, podem ser usadas como fontes de dados para identificação dos padrões para *site*s *web*.

4.4.1 Levantamento de Requisitos

Ao desenvolver um projeto de *Data Webhousing*, é importante saber claramente quais as fontes de dados a serem modeladas. Por meio de um levantamento de requisitos, observou-se como está estruturado o *site*. Identificaram-se os *logs* que serão utilizados para pesquisa como fonte de dados não estruturada e, com base nos dados coletados, construiu-se o modelo dimensional. Qualquer servidor *web* mantém os *logs* de todas as requisições de um *browser,* ou seja, de um navegador de Internet, como data, hora e o endereço IP de onde uma requisição está sendo enviada. Dessa forma, os usuários deixam "rastros" enquanto navegam por pelo *site*. E esses “rastros” serão utilizados como a fonte de informação que ajudará a formular os requisitos para o desenvolvimento do projeto.

Kimball (2000f) destaca que é importante saber que, ao implementar um sistema cuja finalidade é a análise, o refinamento dos requisitos se faz necessário para que se possa ter um resultado o mais apurado possível. O ponto de partida para as demais etapas deste projeto são os passos que serão descritos neste capítulo e que deverão ser revistos à medida que o protótipo for apresentado. Contudo, devido à própria natureza exploratória do trabalho de pesquisa e da própria natureza dos sistemas que se procura atender no ambiente *DW*, em um primeiro momento, não se obtém uma especificação completa de todos os requisitos. Esta especificação se forma com o desenrolar da implementação e validação do protótipo.

O que se sabe, porém, no início são algumas informações que se deseja obter no projeto, ou seja:

- o que mais é acessado no *site*?
- quem acessa o *site*?
- qual é a relação entre os objetivos de um *site* de grupo de P&D e esses acessos?

Por meio de reuniões com os pesquisadores, gestores e especialistas em *Data Warehouse* e *Data Webhouse* foram levantados os requisitos acima. Dessa forma, foram também detectadas outras necessidades de informações, considerando-se diversas situações descritas na seção de análise dos resultados e nos anexos dessa dissertação.

4.4.2 O Projeto

O planejamento da construção do *Data Webhousing* ocorre nesta fase, que se apresenta como uma das mais importantes, pois qualquer falha na delimitação de escopo, identificação de necessidades ou erro na especificação dos recursos pode resultar na inviabilização total do projeto (INMON, 1997).

Em geral, o ponto de partida do projeto é responder a algumas questões de suma importância, tanto para a organização do projeto quanto para os resultados esperados por parte da organização. As perguntas formulas foram feitas no item levantamento de requisitos.

As respostas a estas perguntas ajudam a identificar, claramente, as expectativas e o que se deseja alcançar com essa aplicação e quais as ferramentas serão utilizada para a extração destas informações. Nesta fase, é importante identificar que usuários diferentes possuem objetivos e necessidades diferentes (INMON, 1999).

Com base no perfil de aceitação tecnológica e nível de exigência da organização e de seus usuários, uma estratégia tecnológica pode ser traçada. Nesta fase, pode-se optar pela adoção de um *Data Mart, ad hoc ou Data Webhouse*.

Nessa aplicação, optou-se por um *Data Mart*, que segundo (Barbosa *et., al*. 2002), destaca que não existe uma receita pronta para a construção de um *Data Mart*, o que existe são ferramentas e técnicas que contemplam as várias etapas de um projeto, desde os processos de extração, transformação e análise dos dados coletados até a geração do modelo dimensional no Bando de Dados *Webhousing*. Na seqüência, serão levantadas as principais fases que foram percorridas para a concepção do protótipo.

A Figura 15 demonstra as 3 principais etapas percorridas no desenvolvimento desse projeto, como segue:

- geração dos dados operacionais;
- preparação dos dados na Área de estagiamento;
- apresentação dos dados no *Data Webhousing* por uma ferramenta *Olap.*

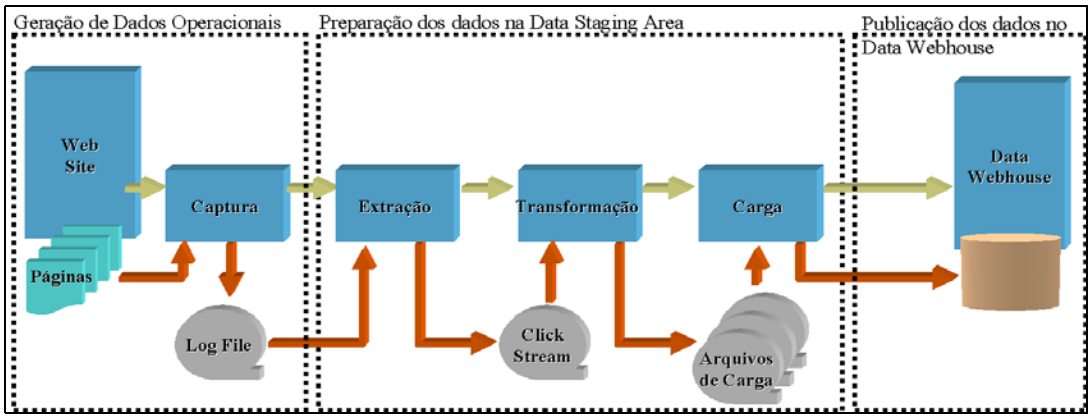


FIGURA 15 - ETAPAS DE IMPLEMENTAÇÃO DO DATA WEBHOUSE - FONTE ADAPTADO DE FARIAS(2002)

A Figura 15 ilustra uma visão dos passos a serem seguidos na implementação do *Data Mart* proposto nesta dissertação.

- Fase 1 - Geração dos Dados Operacionais - cada *site* a ser visitado gera *logs* em arquivo no servidor da Internet da organização onde o *site* está hospedado. Esses arquivos de *logs*, segundo Abiteboul (1997), são denominados dados semi-estruturados, porque diversas fontes de dados semi-estruturados são de grande interesse para várias aplicações, pois apresentam como principais características

o fato de serem ricas em dados e terem abrangência semântica bastante específica, ou seja, os dados disponíveis versam sobre assuntos bem definidos;

A captura desses dados é uma das etapas mais importante no desenvolvimento de um *Data Webhouse*, pois envolve a configuração do servidor e o constante monitoramento pela movimentação dos dados que se vão acumulando nos arquivos definidos para capturar os *logs*.

Para tanto, são extraídos dos arquivos de *logs* do servidor, informações sobre as atividades dos usuários que visitam o *site*, as quais, posteriormente, são transformadas em seqüências de clique. A principal tarefa dessa etapa consiste em armazenar os dados de seqüências de cliques em uma estrutura, onde ficam retidos os dados que, posteriormente, serão carregados para a tabela de estagiamento.

Essa tabela de estagiamento foi desenvolvida na linguagem Progress 7.0, pois a linguagem facilitou a implementação desta fase. A figura 16 apresenta a ferramenta que foi utilizada nessa primeira etapa, onde desenvolveu as rotinas que transportaram os dados não estruturados do servidor para a ferramenta progress 7.0, para serem trabalhadas para a próxima etapa.

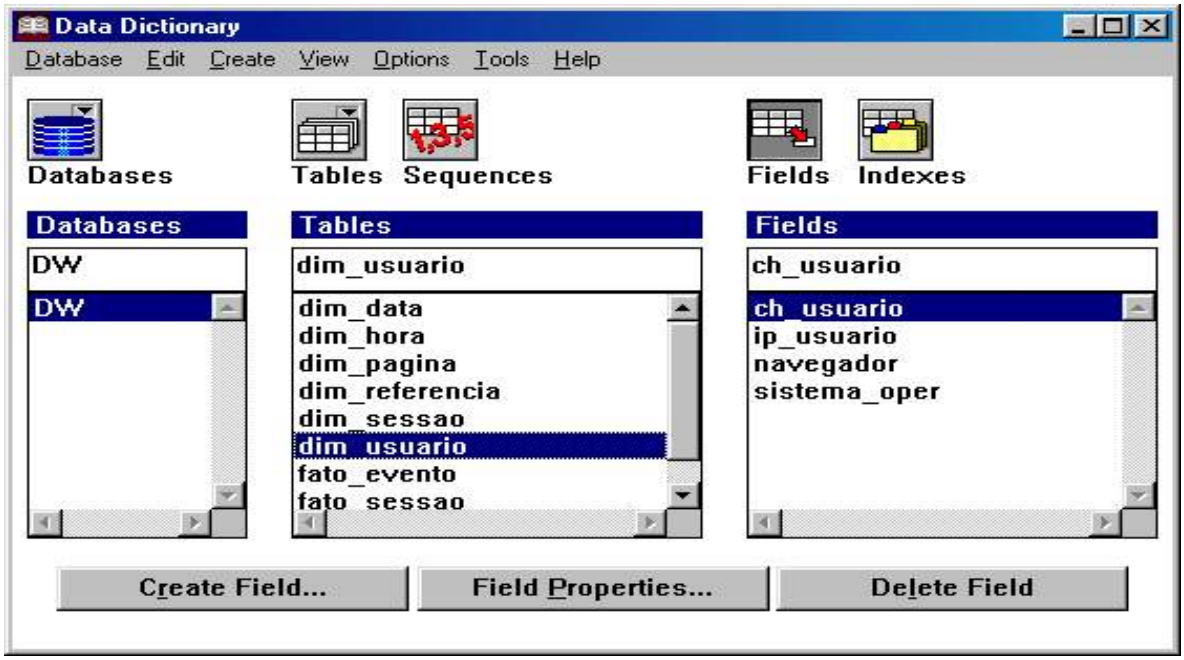


FIGURA 16 – DEFINIÇÃO DO BANCO DE DADOS INTERFACE PROGRESS

Finalizando esta primeira etapa, foi definido um banco de dados temporário, em linguagem Progress, com o objetivo de gravar os dados originados a partir dos arquivos de logs. Este banco de dados consiste em 7 tabelas, sendo 5 tabelas dimensões e 2 tabelas de fatos. Estas tabelas, junto com as definições de campos e chaves, foram construídas via Data Dictionary do Progress.

- Fase 2 – Preparação dos Dados na *Data Staging Area,* (Área de Estagiamento) nesta área preparam-se os dados recebidos da fase de captura de *logs* para uma tabela auxiliar. De posse desses dados, trabalha-se com *software* apropriado, eliminando as informações desnecessárias e preparando os dados para carga no *Data Webhouse*;

Ao desenvolver essas rotinas, foram aplicadas nos arquivos de *logs* onde foram corrigidas as distorções que havia nos dados, por exemplo, duplicação de linhas, *banners* e imagens gravadas, seqüencialmente.

Dessa forma, foram criadas as tabelas numa área auxiliar onde, posteriormente, essas tabelas foram geradas para o modelo dimensional no banco de dados Access 3.0.

Facilitando, assim, a carga dos dados e futuras análises, uma vez que a ferramenta utilizada foi o Excel 9.0, por ser compatível com o modelo de dados apresentado. Esse componente chamado Procedure Editor foi utilizado para automatizar essas rotinas para auxiliar nas extrações dos dados dos arquivos de logs. A Figura 17 apresenta melhor os detalhes dessa rotina.

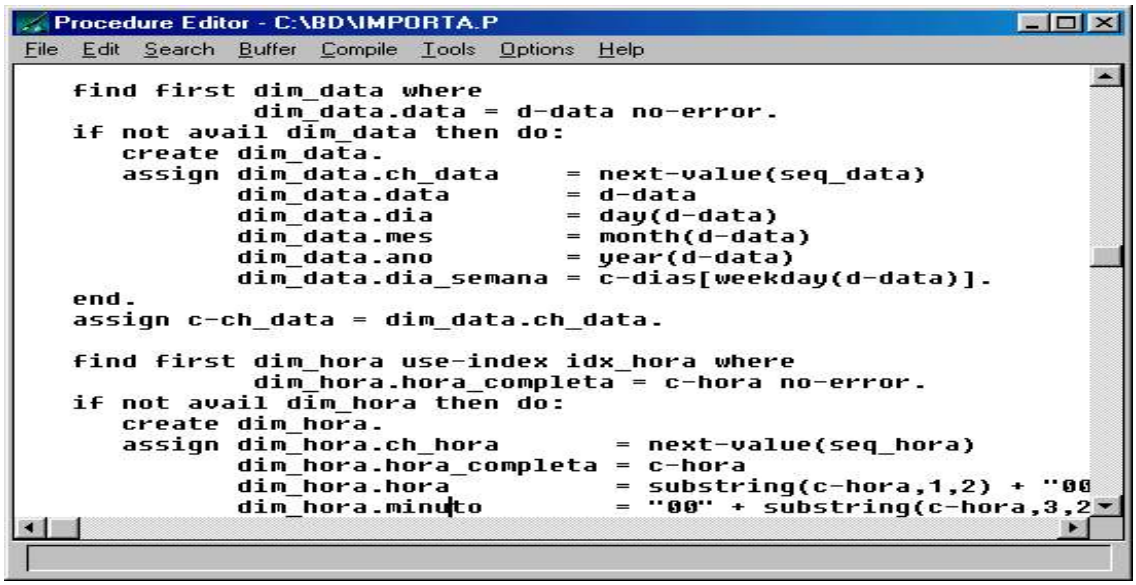


FIGURA 17 – VISÃO GERAL DA ROTINA CONSTRUÍDA PARA FAZER A CARGA DOS DADOS

Com base nas informações acima, as tabelas contidas na base temporária são transportadas para a base Access 3.0, com o objetivo de elaborar as consultas estatísticas nos levantamentos de dados e requisitos deste projeto, em vista disso, as possíveis análises deverão ser feitas utilizando-se a ferramenta Microsoft Excel 9.0. A próxima figura dá uma visão geral da interface do banco Progress 7.0, na qual é constatada a finalização das exportações dos dados tratados para o banco de dados Access 9.0.

A figura 18 apresenta uma visão geral da interface do banco de dados Progress, na qual é visualizado o término das exportações dos dados para o banco de dados Access 3.0.

- Fase 3 - Publicação dos Dados no *Data Webhouse*, é o local onde se importa a tabela da área de estágio para o modelo dimensional já definido no projeto, havendo interação entre esses passos com o ambiente externo, cuja finalidade é a análise das consultas com uma ferramenta *OLAP*.

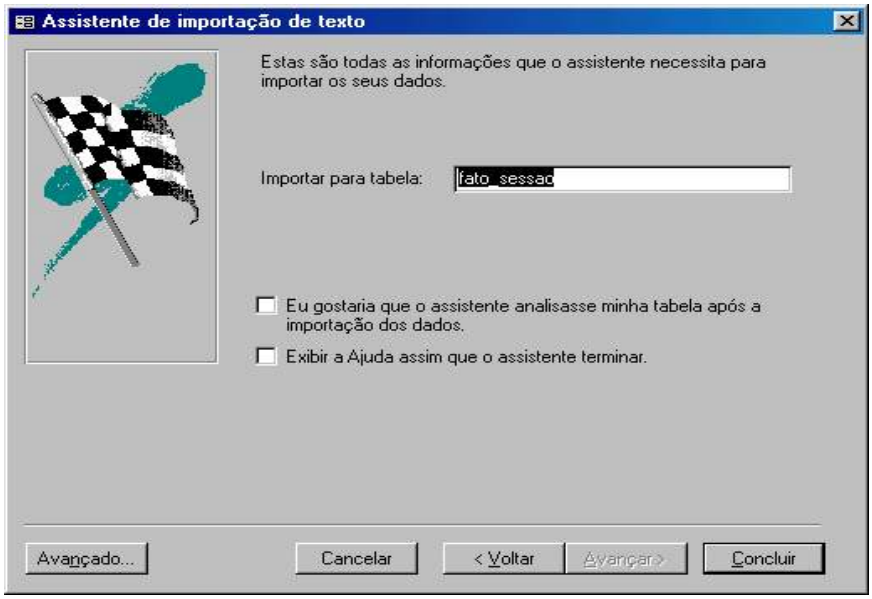


FIGURA 18 – FINALIZAÇÃO DO PROCESSO ETL(EXTRAÇÃO TRANSFORMAÇÃO E CARGA)

O processo de carga é formado por alguns procedimentos nas estruturas intermediárias necessárias para popular o *DM* do projeto proposto. Uma vez validado o esquema estrela, foi implementado o processo de carga já automatizado pelas rotinas de Progress 7.0, e para viabilizar a carga no *DM* foi preciso apenas converter as tabelas da área de estagiamento para o banco Access 3.0.

Tendo concluído todas as etapas de acordo com a figura 17, tem-se uma estrutura relacional normalizada, cujos dados, apesar de terem boa qualidade devido ao tratamento

sofrido, ainda não estão otimizados para consulta, ou seja, ainda não estão prontos para serem usados por uma ferramenta *OLAP*. Para tal, são modelados os *Data Mart* e os processos de extração do *Data Warehouse* para os *Data Marts*, isto é, o modo por meio do qual o modelo dimensional será carregado tomando por base o sistema relacional. Nesse momento, já se deve ter escolhida a ferramenta *OLAP* que será utilizada, uma vez que isso pode afetar a modelagem a ser aplicada no *Data Mart*. A figura 18 apresenta o modelo dimensional gerado já no Banco de Dados.

Esta etapa então finalizou o processo de importação dos dados para o Banco de Dados Microsoft Access, informações contidas dos arquivos textos gerados pela rotina citada na 2º etapa. A próxima etapa será fazer as consultas via cubo de dados, através de uma ferramenta *Olap* ou seja no caso dessa aplicação será utilizada a ferramenta Microsoft Excel.

4.4.3 Ferramenta Olap

A última etapa do projeto contempla as consultas, procurando responder às perguntas para o qual o modelo dimensional foi projetado. Para esse processo, foi utilizada a ferramenta OLAP Excel 9.0 da Microsoft. O termo *OLAP – On-line Analytical Processing* – refere-se a um conjunto de tecnologias voltadas para acesso e análise de dados. Assim, o objetivo final de uma ferramenta *OLAP* é a transformação dos dados em informações capazes de dar suporte a decisões gerenciais de forma amigável e flexível ao usuário e em tempo hábil.

4.4.4 Modelagem Dimensional

Como já foi abordado, o modelo dimensional é composto por tabelas com chaves compostas, denominadas tabelas de fatos, e por um conjunto de tabelas menores, conhecidas como tabelas de dimensões, que possuem chaves simples (formadas por uma única coluna). Em um sentido mais amplo, a chave das tabelas de fatos é uma combinação das chaves das tabelas de dimensão. Isso faz a representação gráfica do modelo dimensional assemelhar-se a uma estrela, e essa é também a razão do modelo ser conhecido como esquema estrela (KIMBALL, 1998).

O Quadro 6 apresenta os elementos do modelo multidimensional do esquema estrela da aplicação que está sendo apresentado para os *sites* de Grupos de P&D. A relação entre esses elementos é representada segundo o modelo dimensional da Figura 19.

Quadro 6 - Elemento do Modelo Dimensional Desenvolvido.

Elemento	Conceito	No Protótipo
Tabela de Fatos	Tabela primária do modelo dimensional, onde cada fato representa uma medida de negócio da organização (Kimball et. al. 1998).	Dados quantitativos sobre o processo de sessão e evento. • Sessão • Evento
Tabela de Dimensão	Tabelas secundárias do modelo dimensional, que guardam um conjunto de tabela relacionadas a uma tabela de fato (Kimball *et. al*. 1998).	Dados relativos a registros de: • Data • Hora • Página • Sessão • Referência Usuário

O modelo dimensional na figura 19 foi definido segundo os critérios adotados nos estudos metodológicos de *Data Webhousing* apresentado no capítulo 2 deste trabalho. Desta forma, a aplicação foi construída de forma a permitir a realização de mudanças que se façam necessárias sem causar impacto sobre a toda a estrutura.

O modelo dimensional, além de agilizar o processamento das consultas, permite uma melhor visualização dos dados, devido à forma simples de organizá-los. Esta forma de organizar os dados ainda propicia a flexibilidade necessária para eventuais ajustes que se façam necessários.

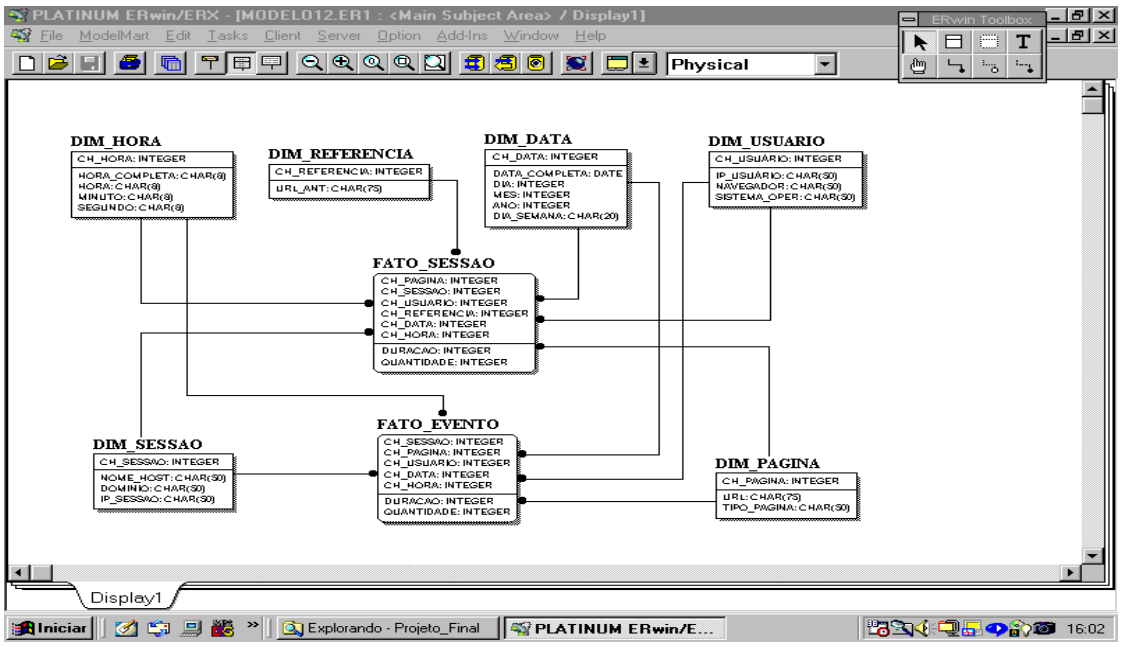


FIGURA 19 - ESQUEMA ESTRELA DESENVOLVIDO PARA A IMPLEMENTAÇÃO DO PROJETO

Após a construção da modelagem foi transformado para o modelo físico, onde foram incluídas características físicas e chaves. No entanto, o esquema-estrela da aplicação foi construído com duas tabelas de fatos e seis tabelas de dimensões, apresentado no Diagrama de Entidade Relacionamento da figura 20 .

De acordo com a figura 19, apresenta duas tabelas principais, que se denominam tabelas de fatos, que contêm os dados do negócio a serem analisados. Em volta das tabelas de fatos, apresentam-se, as seis tabelas descritivas, chamadas tabelas de dimensão. Cada tabela de dimensão possui uma única ligação com as tabelas de fatos, a qual é feita através de chaves externas. O quadro 7 apresenta todos os atributos e descrições apresentadas no modelo dimensional da figura 19 a saber:

Quadro 7 – modelo detalhado dos elementos do esquema estrela

QUADROS	COLUNAS	TIPO DE DADOS	DESCRIÇÃO
fato_sessão	ch_página, ch_sessao, ch_usuario, ch_referencia, ch_data, ch_hora.	Integer Integer Integer Integer Integer integer	Número da página Número da sessão Número do usuário Referência do usuário Data da sessão Hora da sessão
fato_evento	ch_pagina, ch_sessao, ch_usuário, ch_data, ch_hora.	Integer Integer Integer Integer integer	Número da página Número da sessão Número do usuário Data da sessão Hora da sessão
dim_data	ch_data, data_completa, dia, mês, ano, dia_semana.	Integer Date Integer Integer Integer Char(20)	Campo chave data Data completa Dia Mês Ano Dia da semana
dim_hora	ch_hora, hora_completa, hora, minutos, segundos.	Integer Char(8) Char(8) Char(8) Char(8)	Campo chave hora Hora completa Hora local Minuto Segundos
dim_página	ch_página url tipos_página;	Integer Char(75) Char(90)	Campo chave da página Endereço da página Tipo de página
dim_usuário	ch_usuário ip_usuário, navegador, *site*ma_oper.	Integer Char(30) Char(30) Char(30)	Campo chave do usuário Número do ip do usuário Nome do navegador Nome do S.O
dim_referência	ch_referência, url_ant.	Integer Char(75)	Chave da referência Página anterior
dim_sessão	ch_sessão, nome_host, domínio, ip_sessão.	Integer Char(50) Char(50) Char(50)	Chave da sessão Nome da página Domínio Número do ip da sessão

4.4.5 Implementação

Na implementação do projeto *Data Mart*, levou-se em consideração além dos aspectos relativos ao tipo de informação mas também os requisitos exigidos de confiabilidade, rapidez e segurança. Assim, foi definido o banco de dados que armazenou as informações para publicação dos resultados. Contudo, essa combinação de informação e tecnologia aparece nos sistemas de informações, que devem fazer o melhor uso da tecnologia disponível para garantir que os serviços e necessidades dos usuários sejam garantidos.

Dessa forma, os requisitos com usuários e gestores foram concluídos e o banco de dados definido, que deverá ser carregado para o modelo dimensional desenvolvido. No tópico seguinte, deverão ser abordadas as características das ferramentas que foram usadas na implementação do *Data Mart* e o modo como foram utilizadas.

Um aspecto relevante relatado nesta seção é a utilização do banco de dados Access 3.0. Esse banco de dados, além de permitir a projeção do ciclo de vida do *Data Mart*, possibilita a projeção desse ciclo desde a sua concepção. Contudo, é de fácil aprendizagem, pois essa ferramenta possibilita um "ganho de tempo", para fazer as análises foi utilizada a ferramenta Excel 9.0, que é totalmente compatível com o Access 3.0, pois ambos são da família Microsoft, possibilitando assim um ganho de tempo considerável. A Figura 20 exemplifica um diagrama E-R obtido nessa fase, na qual foram relacionadas todas as tabelas de dimensão com as tabelas de fatos.

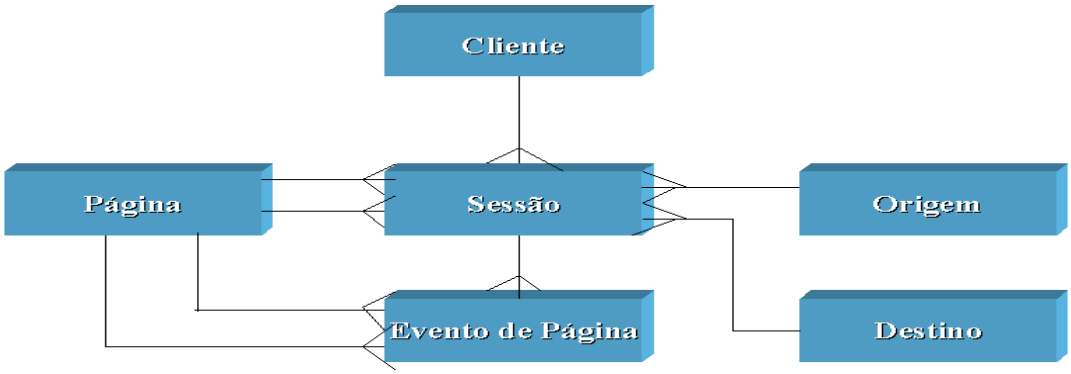


FIGURA 20 – DIAGRAMA ENTIDADE RELACIONAMENTO

É necessário observar a semelhança entre o diagrama E-R da figura 20 e o esquema estrela da Figura 19, apresentada nas páginas acima deste trabalho. Isso indica que o levantamento dos requisitos e o modelo multidimensional são o ponto de partida da especificação do DM, e que a representação E-R é o resultado da visualização desse modelo.

4.4.5.1 Criação das Tabelas no Banco de Dados

Assim, concluiu-se o projeto lógico e físico, devendo ser criadas as instâncias do *DM* no Sistema Geral de Banco de Dados (SGBD). Para tanto, utilizou-se a função específica da ferramenta Erwin, que permite executar os scripts diretamente no esquema específico no servidor de banco de dados.

4.4.6 Ferramentas de Implementação

Para a implementação do protótipo, foram utilizadas as seguintes ferramentas:

- hardware utilizado, computador Atlon 1.1;
- ferramenta de modelagem, software ErWin 3.5.2;
- banco de dados: Access 3.0 e Progress 7.0;
- ferramenta de ETL: Access 3.0 e Progress 7.0;
- ferramenta front-end – Access 3.0 e Excel 9.0.

Há várias razões pela escolha dessas ferramentas, entre elas:

- padronização da Plataforma Windows, uma vez que foram utilizados o Access 3.0 e o Excel 5.0;
- facilidade e disponibilidade de uso, já que a Microsoft disponibiliza essas ferramentas, possibilitando aos usuários conhecerem-nas e a utilizarem-nas, gratuitamente.

4.4.7 Projeto Físico Implementado

Para modelar o *Data Mart,* usou-se a ferramenta case *ErWin* 3.5.2, aplicativo que permite projetar parte do ciclo de vida do *Data Mart* e também possibilita projetar o *Data Mart* desde sua concepção inicial. Todavia, devido à sua funcionalidade, que é própria para modelar *Data Warehouse*, essa ferramenta permite ainda a criação do esquema lógico e físico para o *Data Mart*, permitindo assim um ganho no tempo de desenvolvimento considerável no desenvolvimento do projeto.

Dessa forma, o modelo de dados foi criado no formato esquema estrela com a ferramenta *ErWin* 3.5.2. Essa ferramenta permite visualizar a definição dos nomes das

tabelas, seus atributos e as propriedades de cada um deles. Posteriormente, foram especificadas as relações existentes entre essas tabelas e a respectiva cardinalidade. Por último, foi gerado o modelo dimensional direto no banco de dados. A Figura 21 apresenta o modelo físico do protótipo.

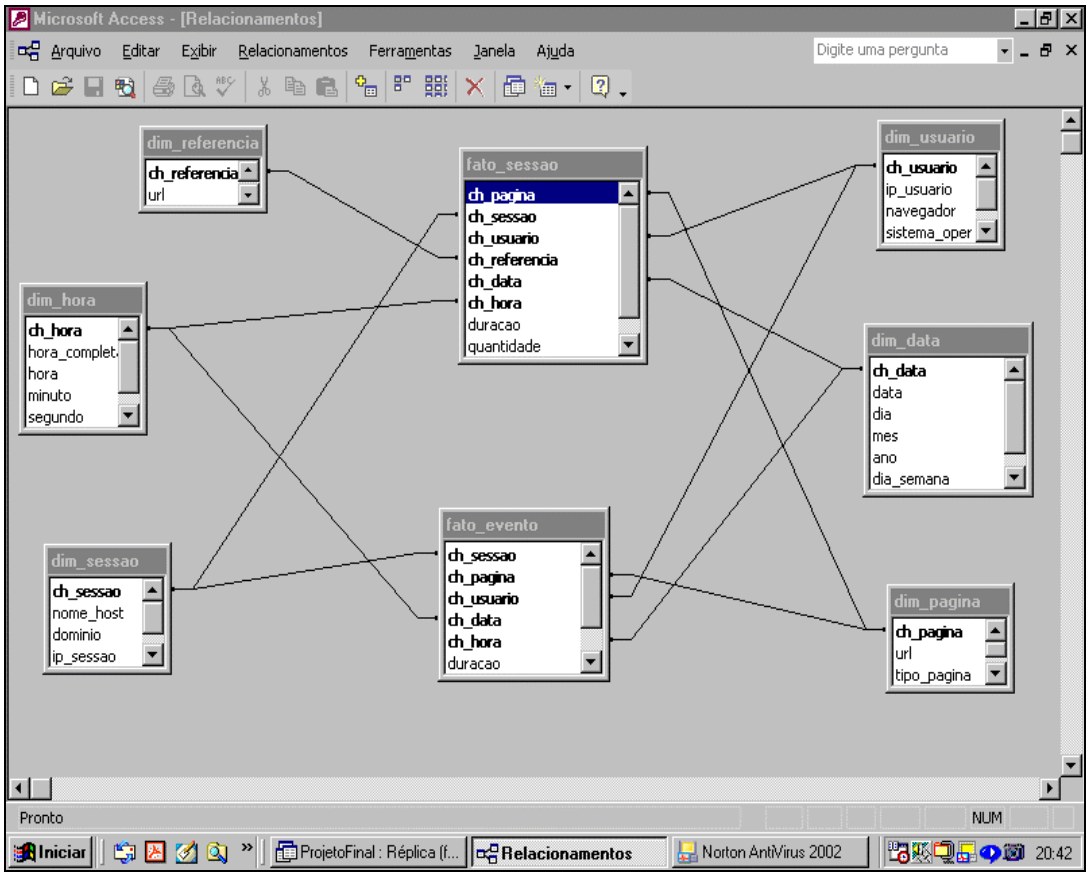


FIGURA 21 – MODELO FÍSICO GERADO NO BANCO DE DADOS

Então, após a confecção das rotinas, elas foram aplicadas nos arquivos de *logs* onde foram corrigidas as distorções que havia nos dados, por exemplo, duplicação de linhas, *banners* e imagens gravadas, seqüencialmente. Finalmente, foram criadas as tabelas numa área auxiliar, onde posteriormente essas mesmas foram geradas para o banco de dados Access 3.0. Dessa forma, as consultas via cubo de dados foram facilitadas, uma vez que a ferramenta utilizada foi o Excel 9.0 por ser totalmente compatível com o modelo de dados apresentado.

As análises dos resultados juntamente com os requisitos levantados nesse capítulo, serão esplanadas, no próximo capítulo.

4.5 Considerações Finais

Para que o desenvolvimento de sistemas de informações acompanhe as mudanças de necessidades dos tomadores de decisões, é necessário que se utilize uma metodologia de desenvolvimento voltada à flexibilidade e ao processo evolutivo contínuo, o que leva ao uso dos *Data Warehouses e Webhouses*. Dessa forma, este capítulo apresentou os estudos dos *sites web* de Grupos de P&D, bem como também apresentou os processos de uma aplicação *Data Mart* de seqüências de cliques.

De acordo com os processos analisados, foi implantada uma metodologia, passo a passo, num ambiente *Data Webhouse*. Na proposta implementada, utilizou-se como estudo de caso, um *site* específico de um Grupo de Pesquisa e Desenvolvimento da UFSC, e para esse estudo, observaram-se todas as etapas de um processo de extração, transformação e carga de dados.

Na construção dessa aplicação, foram estudados todos requisitos necessários para auxiliar os resultados estatísticos numa proposta de melhorias para os *site*s de grupos de P&D. Assim sendo, a implantação de um *Data Webhouse*, fornece subsídios relevantes para o processo de melhorias nas estruturas e nos objetivos dos *sites* de P&D.

Conclui-se, portanto, que a modelagem dimensional dispõe de um grande potencial para as aplicações de *Data Webhouse* e que seus elementos apresentam técnicas e passos que orientam o seu desenvolvimento e aumentam a probabilidade de sucesso do projeto. A criação de um *Data Mart,* utilizando seqüências de cliques, fornece informações imprescindíveis para as instituições das mais diversas áreas.

O próximo capítulo apresentará os resultados obtidos e as análises das consultas solicitadas ao modelo dimensional implementado.

5 ANÁLISE DOS RESULTADOS OBTIDOS

5.1 Considerações Iniciais

Conforme foi apresentado, no capítulo anterior, a aplicação de *Data Webhousing* em *sites* de grupo de P&D, visa subsidiar o processo de organização do conteúdo e estrutura de informações.

Este capítulo apresenta os resultados de uma aplicação de *Data Webhousing* sobre os acessos ao *site* de um Grupo de P&D em uma instituição de ensino superior. A partir das observações registradas no desenvolvimento da aplicação e na interpretação dos dados coletados, verificou-se que o grande desafio de uma proposta de configuração de indicadores está na interpretação quanto à sua definição, e não na sua proposição. Assim, aliando-se os aspectos abordados neste capítulo, os resultados apresentados centraram-se na observação de cada indicador, bem como a que se destinam.

A ênfase da proposta para avaliar as estruturas dos *sites* de P&D está na utilização desses indicadores para inferir o perfil padrão de comportamento, interesses dos usuários e adaptar a estrutura do *site* de acordo com as informações analisadas.

Contudo as considerações feitas sobre a análise resultante do protótipo, poderão ser levadas em conta na reestruturação dos *sites* de Grupos de Pesquisa e Desenvolvimento e devem orientar futuras reestruturações sobre o mesmo *site*, dado o caráter contínuo de análise e reestruturação proposta neste capítulo.

Cabe salientar, ainda, que será proposto na última seção deste trabalho um modelo referencial de *sites*, em caráter ilustrativo para os Grupos de P&D.

5.2 Análise dos Resultados Via Cubo de Dados

Conforme foi apresentado, no capítulo 2, sobre a perspectiva de utilização da abordagem dimensional para representação dos dados, o modelo dimensional lembra a idéia do cubo (Cielo, 2001), contendo três ou mais dimensões, cada um representando um atributo diferente, conforme apresenta a Figura 22. O modelo dimensional, além de agilizar o processamento das consultas, permite uma melhor visualização dos dados, devido à forma simples de organizá-los. Esta forma de organizar os dados ainda propicia a flexibilidade

necessária para eventuais ajustes que se façam necessários no modelo. Dessa forma, é comum se associar a tecnologia *OLAP* à manipulação multidimensional dos dados. Estas estruturas de dados permitem que os dados sejam apresentados e analisados sob a ótica do gerente ou do tomador de decisão, facilitando a análise de dados, através de sumarizações[5], de acordo com a definição do modelo, aliando esta análise à possibilidade de visualizar qualquer intervalo de tempo definido no *Data Webhousing.*

Dessa forma, utilizou-se a planilha Excel 9.0 como ferramenta *Olap* para suporte e análise dos dados coletados. Esta ferramenta permite a criação de cubos a partir de uma fonte de dados. As consultas relacionadas foram formuladas no levantamento de requisitos do projeto.

Como exemplo, tomam-se as dimensões de página e usuário, onde se pode localizar determinado fato, por exemplo, que páginas são mais acessadas no *site* ou quais são os endereços que acessam esse *site.* A partir do cruzamento dos dados relacionando-se as dimensões às tabelas de fatos pertinentes, o usuário "gira" o cubo para obter novos fatos baseando-se nas dimensões definidas. A figura 22, simboliza um cubo multidimencional.

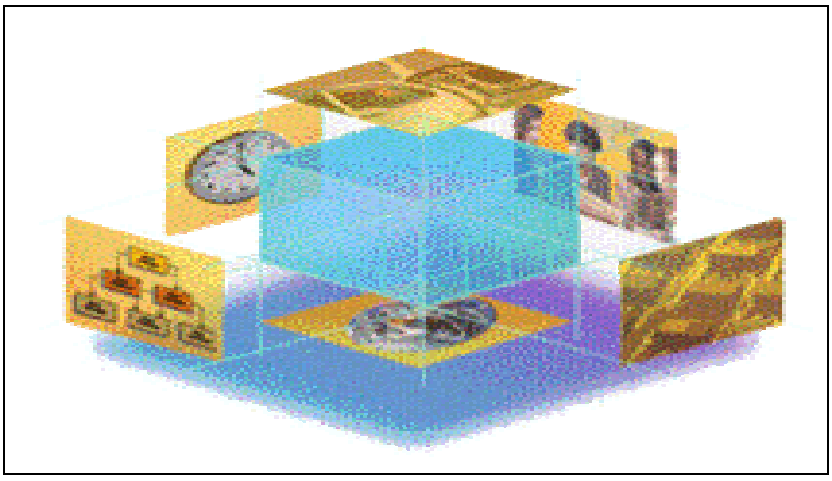


FIGURA 22 - SEMELHANÇA DO MODELO DIMENSIONAL A UM CUBO – FONTE CIELO (2001)

Dessa maneira, as estatísticas que serão mostradas são de um Grupo de Pesquisa e Desenvolvimento da UFSC, que teve como objetivo registrar os acessos feitos a este *site*. O período de captura dos dados compreende as datas de 6/8/2002 a 16/8/2002, ou seja, um período de dez dias de monitoramento. Esse período foi determinado a fim de limitar-se a

[5] Sumarização é a transformação de informações de um nível mais baixo de granularidade para um nível mais alto. Isto evita a redundância e o desperdício de tempo com os cálculos mais comuns.

quantidade de dados a serem manipulados pelos aplicativos de extração, sem comprometer, entretanto, a qualidade das análises solicitadas no levantamento de requisitos.

O processo de *Data Webhousing* conduz o gestor do *site* a descobrir os perfis e os padrões de acesso dos usuários que estão acessando *sites na web.* Como os Grupos de P&D disponibilizam informações em seus *site*s de forma desestruturada, a coleta de dados na *web* e análise dos resultados, além de obter respostas para as questões que orientam o gestor na tomada de decisão, também conduz a propiciar uma melhor reflexão sobre o processo de reestruturação do seu *site*.

Dessa forma, diferentes *sites* de Grupos de P&D podem disponibilizar suas informações, bem como a estruturá-los, de acordo com os objetivos levantados neste projeto.

A necessidade declarada por informações estratégicas para inserirem em seus *sites,* constatou-se que os pesquisadores não dispõem de um mecanismo adequado para tratar os dados gerados em seus *sites*. A questão está no fato de o volume de dados ser muito grande se ele não estiver organizado de forma a agregar valores, o que significa transformar dados em informações úteis. Como foi visto na revisão da literatura, no capítulo 3, a modelagem dimensional é uma abordagem que através de ferramentas *Olap,* pode dar sentido a esta montanha de dados. O modelo dimensional representa os dados como matriz, na qual cada dimensão é um tema ou assunto de negócio que será objeto da análise e o tempo é sempre uma das dimensões consideradas.

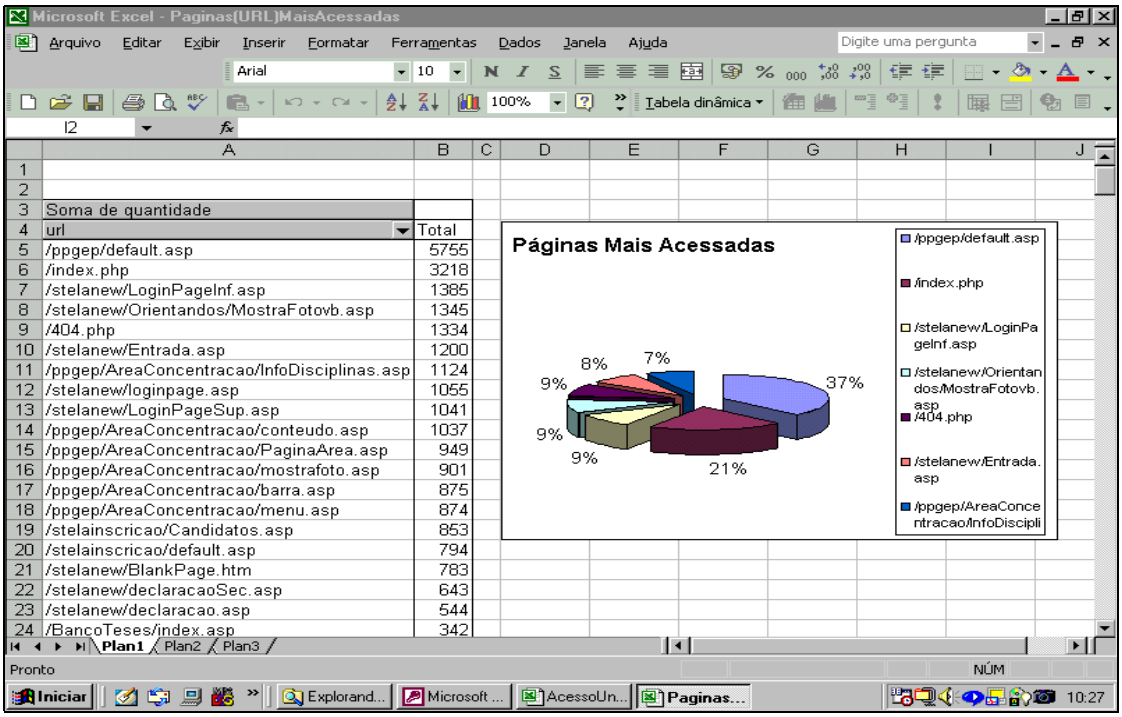
Nesse tipo de ferramenta, o usuário pode solicitar as mais variadas consultas, de acordo com as sumarizações desejadas, ficando a cargo do gestor a definição de cada uma das visões que melhor se adequar à resposta esperada.

5.3 O que é mais acessado no site?

De acordo com o levantamento de requisitos, descrito na seção 4.4 do capítulo 4, a primeira pergunta a ser respondida no protótipo configurou-se da seguinte forma: Foram solicitadas quais as páginas mais visitadas no *site*? Dessa forma, foram selecionadas as tabelas DIM_PÁGINA e FATO_SESSÃO. Com a combinação dessas tabelas, utilizando-se as técnicas de *Data Warehouse*, então faz a consulta de acordo com o fato determinado. Com isso aplicou-se as técnicas *Olap,* ou seja, configura-se a consulta com a ferramenta Excel 9.0

da Microsoft, através da opção Cubo de Dados para obter as informações requisitadas do modelo.

Dessa forma, o modelo dimensional selecionou o atributo chave ch_página, na tabela fato_sessão. A Figura 23 ilustra os resultados estatísticos referentes à análise das páginas mais acessadas do *site* em estudo.



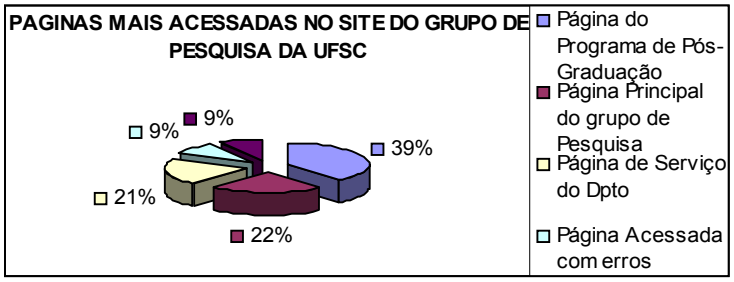
Soma de quantidade	
url	Total
/ppgep/default.asp	5755
/index.php	3218
/stelanew/LoginPageInf.asp	1385
/stelanew/Orientandos/MostraFotovb.asp	1345
/404.php	1334
/stelanew/Entrada.asp	1200
/ppgep/AreaConcentracao/InfoDisciplinas.asp	1124
/stelanew/loginpage.asp	1055
/stelanew/LoginPageSup.asp	1041
/ppgep/AreaConcentracao/conteudo.asp	1037
/ppgep/AreaConcentracao/PaginaArea.asp	949
/ppgep/AreaConcentracao/mostrafoto.asp	901
/ppgep/AreaConcentracao/barra.asp	875
/ppgep/AreaConcentracao/menu.asp	874
/stelainscricao/Candidatos.asp	853
/stelainscricao/default.asp	794
/stelanew/BlankPage.htm	783
/stelanew/declaracaoSec.asp	643
/stelanew/declaracao.asp	544
/BancoTeses/index.asp	342

FIGURA 23 – PÁGINAS MAIS ACESSADAS NO *SITE*

Ao registrar o resultado desta primeira análise, observou-se que as páginas mais visitadas pelos usuários, conforme a Figura 23, são os serviços prestados pelo Grupo de Pesquisa aos alunos do PPGEP. Dessa forma, esta análise indica que 37% dos visitantes, acessam o *site* em busca de informações ao Programa de Pós-Graduação, mantido no *site* do Grupo de Pesquisa analisado, além de outros serviços disponibilizados pelo Grupo ao Programa de Pós-Graduação, como: os Serviços Stelanet, Banco de Teses e Dissertações e a Plataforma de Currículos Lattes.

A Figura 24 apresenta um resumo dos principais acessos agrupando as páginas campeãs de acesso ao *site*.

Com base nessas informações, os mantenedores do *site* poderão avaliar os serviços e tomar decisões de como melhorar os serviços do *site*, levando em consideração, o objetivo dos Grupos de Pesquisa, que é mostrar a sua produção de C&T.



URL DO SITE	Nª de ACESSOS
Página do Programa de Pós-Graduação	5.755
Página Principal do grupo de Pesquisa	3.218
Página de Serviços do Depto	3.032
Página Acessada com Erros	1.334
Páginas de Projetos Desenvolvidos	1.267

FIGURA 24 – VISÃO GERAL DOS ACESSOS AGRUPADOS

O agrupamento configurou-se da seguinte forma: PPGEP obteve 5.755 visitas com 43% dos acessos. As páginas de serviços foram agrupadas como: Stelanew, página para inscrição de candidatos ao PPGEP. StelaNet e StelaMining com 3.218 visitas, que são as páginas principais do *site* e corresponde a uma porcentagem de 24% do total de acessos. O Banco de Teses/Dissertações (Páginas de Serviços de Depto), e a Plataforma Lattes, (Projetos do Grupo) corresponderam a 4.299 visitas, num total de 34% dos acessos e finalmente com 1.334 acessos, as páginas que não obtiveram sucesso ou seja o usuário tentou acessar e encontrou erro, correspondendo a 9% do total dos acessos.

Tomando-se como base a distribuição dos acessos às páginas da Plataforma Stela, destacam-se mais as atividades relacionadas ao curso de mestrado, doutorado e prestação de serviços de projetos do PPGEP, que atingiram 43% das visitas totais do *site*. Vale ressaltar, ainda, que durante a avaliação do *site* com os indicativos apresentados, nas Figuras 23 e 24, o *site* estudado por possuírem características próprias de desenvolvimento de sistemas de Informação priorizou destacar mais as atividades de prestação de serviços e os projetos desenvolvidos e em desenvolvimento.

5.4 Quem Acessa o Site?

Baseando-se nos levantamento de requisitos descritos na seção 4.4 do capítulo 4, a segunda questão indagada pelos gestores do *site* configurou-se da seguinte forma: Quem são os usuários que mais acessam o site? De acordo com a figura 19 do capítulo anterior, no modelo dimensional, desenvolvido para solucionar as questões solicitadas junto ao levantamento de requisitos, foram selecionadas as tabelas DIM_USUÁRIOS e FATO_SESSÃO, nas quais posteriormente, por meio dos recursos *Olap* da ferramenta Excel 9.0, elaborou-se a consulta através do cubo *olap*, onde se obteveram as informações desejadas definidas nos levantamento de requisitos, vale dizer que quem são os usuários que estão mais acessando o *site?* Então o *Data Mart* selecionou os Ips (Protocolo de Internet) dos computadores que acessaram as Páginas do Grupo de Pesquisa, em estudo, e com os endereços de Ip, obteve-se o endereço dos *sites* das pessoas que estão acessando o *site*. Na Figura 25, visualiza-se a freqüência dos usuários que visitaram as páginas do *site*.

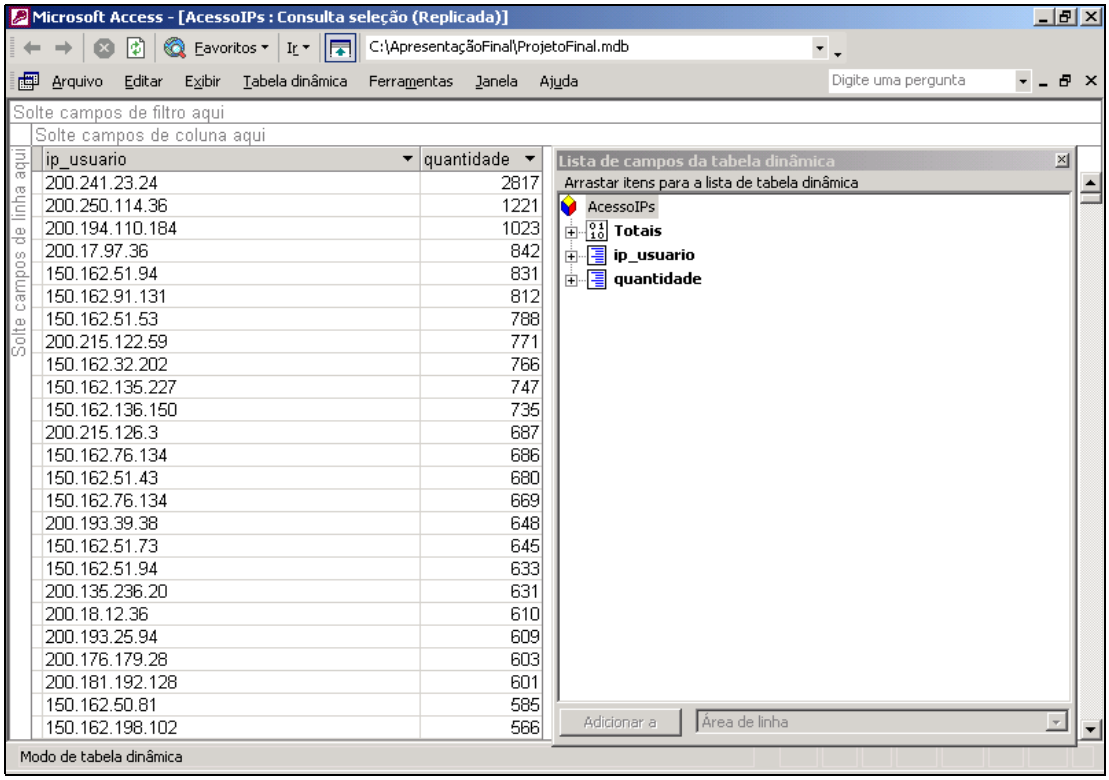


FIGURA 25 – VISÃO GERAL DOS IPS MAIS ACESSADOS NO SITE

Essa visualização se refere aos Ips que mais acessaram o *site*, dessa forma, o endereço portador do Ip 200.241.23.24 acessou o *site* do Grupo de Pesquisa 2.817 vezes, em seguida o Ip 200.250.114.36 acessou 1.221vezes. Nessa visão, essa consulta agrupou todos os Ips em ordem numérica, destacando-se os Ips que mais acessaram e os que menos acessaram o *site*.

Para essa relação de *sites* mais acessados apresentada na figura 25, recomenda-se ao gestor ou administrador do site, que para saber o endereço do domínio dos Ips, que utilize ferramentas adequadas como o (NSLOOKUP), ou, então, sugere-se que desenvolva uma aplicação que leia essa relação de endereços e procure o domínio na *web* a qual endereço o Ip pertence. Assim, o gestor poderá ter descrição de todos os domínios e então poderão ter uma maior interação com os usuários potenciais de seu Grupo de P&D.

5.5 Qual é a Relação Entre os Objetivos do *Site* Analisado e os Acessos Obtidos?

O objetivo do *site* de um Grupo de Pesquisa é a organizar e disponibilizar, de forma sistêmica, as informações sobre a pesquisa institucional, produção científica e sobre o portfólio de produtos desenvolvido pelo Grupo ou seja tornar visível a sua produção cientifica e tecnológica.

Quanto à relação dos objetivos do *site* e seus acessos, observou-se uma premente necessidade de reestruturação, pois a maior parte dos acessos feitos sobre o *site* analisado estavam relacionados ao conteúdo de informação do departamento onde o Grupo está lotado.

Vale justificar, porém que o *site* do Grupo Stela por ter característica própria apresenta recursos de informação sobre a área em que se desenvolve; por isso, o *site* de Grupo de Pesquisa é uma Plataforma de Informações, onde apresenta o Grupo de Pesquisa e todos os seus principais projetos de Gestão em C&T. Por isso, que indicativos de maior números de acessos são de serviços e Projetos, Projetos esses de grande relevância em nível nacional como o Banco de Teses e Dissertações e o Projeto da Plataforma de Currículos Lattes do CNPq. Com isso vale dizer que um Grupo de Pesquisa que tem característica própria, recurso de informação sobre sua área fim e sobre tudo quando tem sistemas de informação, deve procurar separar seu *site* institucional dos *sites* com recursos informacionais em links oferecidos pelo Grupo de Pesquisa.

Dessa forma, o método de avaliação proposto por este trabalho leva em consideração os objetivos de Grupos de Pesquisa e Desenvolvimento definido e apresentado no capítulo 4, seção 4.2, Quadro 5 desta dissertação.

Levando-se em consideração as análises feitas, neste trabalho, em conjunto com um trabalho extenso de design e diagramação de conteúdo, foi elaborada uma nova proposta para

organização de *sites* de Grupos de P&D, onde foi dado especial destaque sobre os projetos realizados em *links* próprios, as linhas de pesquisa do grupo, ao perfil da equipe, ao histórico do grupo, aos principais clientes, portfólio de produtos e a equipe de trabalhos.

Dessa maneira, o *site* foi organizado de forma a privilegiar clientes e parceiros potenciais para os Grupos de P&D.

5.6 Proposta de Melhorias

Conforme foi ilustrado na Figura 26, definiu-se uma proposta para reestruturação contínua do *site* de acordo com o processo apresentado no capítulo 4, ou seja, a transformação de um *web site* em um *site* adaptativo. Classifica-se como *site* adaptativo, *web sites* que utilizam as informações sobre o padrão de acesso dos seus usuários e considerando a estratégia da organização para melhorar a estrutura das suas páginas e melhor reorganizar o seu conteúdo.

Contudo, a Figura 26 apresenta um processo de melhorias para *sites* de P&D, envolvendo seis fases previamente descritas: levantamento dos objetivos do *site* do grupo analisado, interpretação dos objetivos do *site,* observação e transformação das informações, aplicação das técnicas *Data Webhousing*, diagnóstico das análises de resultados e reestruturação do *site*.

Portanto, executadas essas fases, a estratégia e os objetivos dos usuários para a estruturação dos *sites* estão definidas.

De acordo com a metodologia descrita acima, as Figuras 27 e 28, apresentam uma sugestão de um novo *site,* onde é demonstrado um modelo de *sites* para Grupos de Pesquisa e Desenvolvimento.

Diante dessa proposta, o objetivo será integrar os Grupos de Pesquisa para uma melhor interação e, possivelmente, dar mais qualidade às pesquisas desenvolvida nas IES. Diante dessa proposta, possibilitará ao usuário uma melhor navegabilidade de acordo com os objetivos dos Grupos de Pesquisa. O *site* ilustrativo apresentado está definido de acordo com o quadro 5 no capítulo 4.

Contudo, a contribuição que essa proposta trará para o desenvolvimento de um novo *site* se encontra esquematizada na Figura 26. Uma vez que os dados suficientes já foram coletados e observados, é preciso descobrir que tipos de informações os usuários precisam

inserir no *site*, ou seja, o padrão de comportamento, os perfis dos usuários, para que a estrutura do *site* possa ser remodelada. Os dados coletados na fase de observação precisam ser trabalhados, relacionados e bem entendidos a fim de que informações válidas sejam produzidas.

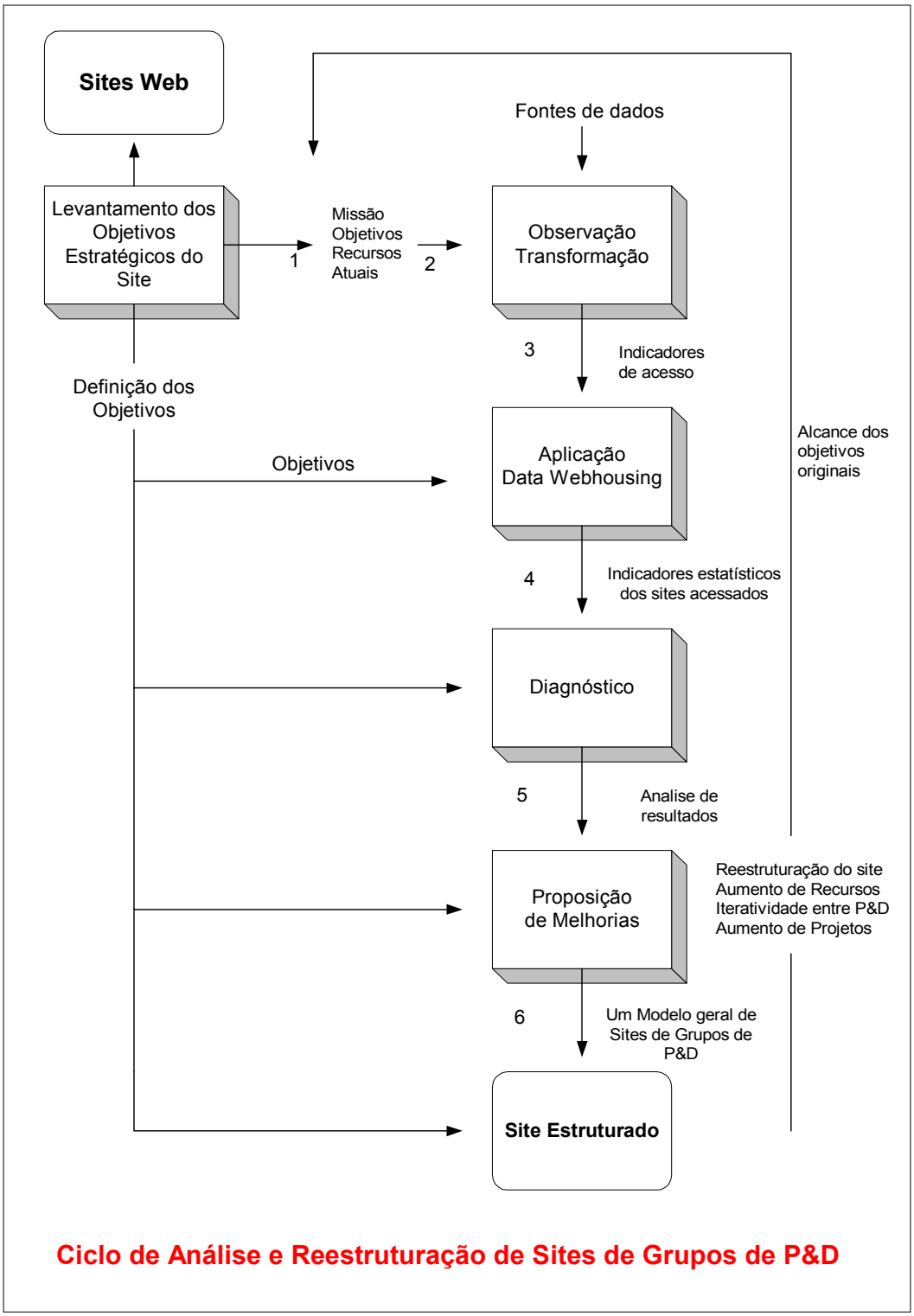


FIGURA 26 – METODOLOGIA APLICADA PARA REESTRUTURAÇÃO DE SITES DE P&D

Dessa forma, a solução proposta é apresentada como uma estrutura dividida em seis etapas, cuja realização se deu da seguinte forma:

- *Site*s de P&D – Nesta etapa apresenta-se o *site* e faz-se um estudo que indique como se deseja que o *site* seja visto na Internet; para tanto, será feita uma análise que leva em consideração os objetivos estratégicos do *site*. Somente com esses objetivos definidos é que as estratégias poderão ser projetadas. O próximo passo do processo deve ser a coleta de dados.

- Na fase de observação, são coletados os dados relacionados às informações dos usuários com o *site*. Deve-se ressaltar que não é possível a realização de uma análise da eficiência da estrutura do site, com o fim de alterar essa estrutura, se não existirem dados sobre a interação dos usuários com o *site*. Os dados observados podem ser caracterizados pelos *logs* dos servidores de Internet, onde as informações de acesso dos usuários podem ser obtidas, sem que o usuário precise responder a qualquer tipo de questionário. Num arquivo de *log,* encontram-se todas as requisições feitas ao servidor *web*, incluindo IP da máquina que fez as requisições e a data/hora do acesso. Dessa forma, é possível identificarem-se sessões e observar-se qual a ordem de acesso das páginas de um *site* durante uma sessão. Com essas informações, há a possibilidade da realização de aplicações para extração desses dados para um modelo mais adequado a operações analíticas, o modelo dimensional (KIMBALL, 1998).

Após a fase de observação, ou seja, depois que os dados já foram coletados, é necessário transformar esses dados de uma forma que a aplicação possa reconhecê-los. As informações coletadas na fase de observação precisam ser trabalhadas, relacionadas e bem entendidas para que a técnica de *Data Webhousing* possam ser aplicadas.

- *Data Webhousing* – De acordo com Kimball (2000f), deverá ser elaborado um *Data Mart* de seqüências de clique sobre o qual se aplicam as técnicas de *Data Webhousing* para observação dos acessos aos *site*s. Nessa fase da proposta de melhorias, o usuário deverá analisar os indicativos estatísticos fornecidos pelo *Data Mart,* os quais fornecerão subsídios para a reestruturação do site analisado.

- Diagnóstico – Nesta fase, o usuário do projeto deverá observar os indicativos estatísticos obtidos na fase anterior, de acordo com o levantamento de requisito.

O *Data Mart* apresentará os seguintes resultados:

- página mais acessada no seu *site;*
- página que teve menos acessos;
- de onde vêm os acessos;
- quem acessa esse *site;*
- caminho utilizado pelo usuário na navegação do *site*.

De posse das informações, o tomador de decisão terá informações seguras para verificar se seu *site* é eficaz quanto aos seus objetivos e a missão do *site*.

- Proposição de Melhorias – Na última fase do projeto, apresentam-se as análises das verificações realizadas e então se confrontam essa análises com os objetivos iniciais, procurando-se observar por meio dos acessos às páginas do *site*, se o seu objetivo foi atingido, de acordo com o modelo geral de *sites* de grupos de P&D.

Com base nas sugestões acima, e no modelo apresentado, na Figura 26, apresenta-se uma forma para representar as estratégias de melhoria para estruturas de *sites web*. Essa modelagem permite ainda as especificações de todas as variáveis de um processo de personalização. Contudo, a proposta identifica-se como uma ferramenta útil ao *Data Webhousing,* por ser uma ferramenta que auxilia na estruturação de *site*s de Grupos de P&D e na interpretação dos dados coletados na pesquisa. Dessa forma, essa metodologia permitiu apresentar uma das principais contribuições onde se destaca o modelo geral de *sites* para Grupos de P&D, apresentado nas figuras 27 e 28.

5.6.1 Modelo Referencial para Sites de Grupos de P&D

Além da elaboração das páginas, das informações e principalmente dos *links* que serão disponibilizados, devem ser adotados procedimentos, principalmente os de organização física do *web site*, ou seja, a organização das páginas. Independente de qual estrutura será utilizada, é fundamental manter uma organização hierárquica da informação disponibilizada, que permita ao usuário do *site* manter-se nele o maior tempo possível.

Toda metodologia de estruturação de *web site* deve buscar a valorização e o incentivo a disponibilização de informações, como meio de divulgação de produtos, serviços e produção cientifica. Garantir uma maior visibilidade a todas as informações disponibilizadas

em qualquer servidor *web*, através do qual será possível o acesso ao conteúdo de toda a informação *online* existente internamente. As etapas de criação, desenvolvimento e disponibilização das informações devem ser um trabalho que permita estabelecer uma filosofia voltada para o usuário. As reavaliações devem ser constantes, e sempre garantir uma maior interatividade e participação dos usuários com o *web site*.

A seguir apresentam-se o *web site* Grupo de Pesquisa e Melhoramento Genético Animal (GP/MGA), como sugestão para a estruturação do conteúdo e ilustrando os itens que devem estar relacionados em um site fr Grupo de P&D. O *site* foi desenvolvido em caráter ilustrativo, lembrando que para o desenvolvimento de *web sites*, é preciso que os profissionais das informações estejam capacitados para se tornarem auto-suficientes no uso das tecnologias da informação disponíveis, e mais do que nunca, estarem preparados para trabalharem em equipes multidisciplinares tanto quanto é multidisciplinar a Internet.

FIGURA 27 – MODELO DE SITES PARA OS GRUPOS DE P&D (LINHAS DE PESQUISA)

FIGURA 28 – DISTRIBUIÇÃO DE CONTEÚDO NO MODELO PARA OS SITES DE GRUPOS DE P&D

Dessa forma, serão apresentadas algumas definições sucintas dos links sugeridos para que os Grupo de Pesquisa possam atingir os objetivos definidos neste trabalhos . As Figuras 27 e 28 apresentam uma visão geral do *site.* De forma geral, o *site* compreende as seguintes informações:

Missão do Grupo – Os Grupos de Pesquisas de uma maneira geral tem por missão a promoção da pesquisa, desenvolvimento, formação e extensão nas áreas em que atuam. Organizam e disponibilizam de forma sistêmica as informações sobre a Pesquisa Institucional e a Produção Científica. Articulam com às agências de fomento, a obtenção de recursos financeiros, viabilizando o desenvolvimento da Iniciação Científica e proporcionando meios ao Comitê de Pesquisa para acompanhar e avaliar, a fim de consolidar o Programa de Pesquisa na Instituição.

Apresentação ou Histórico do Grupo de Pesquisa – Para documentar o Grupo, é necessária uma apresentação histórica de como foi criado o Grupo e as principais atividades que envolvem seus membros com o objetivo de conhecer melhor as características de pesquisa desenvolvida pelos Grupos.

Equipe – Mostrar toda a equipe que compõem o grupo, sua titulação área de pesquisa e projeto envolvido.

Projetos de Pesquisa – Esta página concentrará todos os projetos de pesquisa do Grupo, membros participantes e descrição de todos os projetos e áreas de atuação de cada pesquisador. Descrever nesta página também através de ações institucionais, os projetos de pesquisa básica e aplicada garantindo sua indissociabilidade do ensino e da extensão. Estimular as propostas de ações interdisciplinares e interinstitucionais. Prover centros e núcleos de apoio para o desenvolvimento das atividades de investigação. Induzir projetos de pesquisa socialmente significativos na área da saúde incluindo a área de educação específica.

Produção Científica - Esta página permite que os Gestores responsáveis pelo Grupo tenham o total controle das produções científicas de toda a equipe. Através deste link,o usuário poderá possuir mais alguns *links*. Esta página deverá apresentar a produção científica dos Pesquisadores do GP/MGA, dividida nas seguintes seções:

- Trabalhos publicados em revistas e periódicos nacionais e internacionais
- Trabalhos publicados em congressos
- Teses e dissertações de mestrado e doutorado concluídas
- Teses e dissertações em andamento

Teses & Dissertações – Mostrar as teses e dissertações orientadas pelos pesquisadores do Grupo como também os alunos participantes dos projetos e bolsista de Pibic que trabalham no Grupo.

Linhas de Pesquisa – Nesta página, serão apresentadas às linhas de pesquisa dos componentes do Grupo, como mostra a Figura 30 acima. As linhas de pesquisa que esse site apresenta são:

- Peixes;
- Monogástricos;
- Ruminantes.

Bibliotecas – A biblioteca destina-se a dar apoio a todas as atividades de pesquisa do grupo e da comunidade científica. Uma biblioteca de referência, especializada na área do Grupo de Pesquisa é necessária pois assim facilitará a pesquisa e a extensão de tecnologia ligada às áreas de atuação que o grupo está pesquisando, principalmente, se tiver interagindo com outros grupos da mesma área. Também fazem parte de seus acervos as teses, dissertações de todos colaboradores do Grupo. Assim faz-se necessário indicar referências as bibliotecas onde o Grupo atua como área de pesquisa.

Portifolio - O portfolio (do inglês) é uma modalidade de avaliação que tem o objetivo de criar novas formas de avaliar o desenvolvimento das inteligências artísticas. O seu conceito surgiu na história das artes e denomina um conjunto de trabalhos de um artista (desenhista, cartunista, fotógrafo etc.) ou de fotos de ator ou modelo usado para divulgação das produções entre os clientes. Nesse caso, é um instrumento útil pela possibilidade de poder comprovar os trabalhos individuais exemplares, as suas capacidades criadoras e artísticas. Em algumas profissões como no caso do estilismo, do cinema, da fotografia, da arquitetura, do marketing ou do design, são as próprias características pessoais, institucionais ou de produto de pesquisa que, mais normalmente, se procuram demonstrar com o objetivo de entrada no mercado e, dessa forma é, que os documentos arquivados pretendem trazer à evidência.

5.8 Considerações Finais

O objetivo deste capítulo foi demonstrar as análises e considerações obtidas sobre o protótipo apresentado no capítulo 4, além de propor uma metodologia para melhorias para os *web sites* de Grupos de P&D, também foi proposto um modelo *site* para os Grupos de Pesquisa, que por meio de um estudo de caso prático, explorou-se uma experiência com um Grupo P&D da UFSC.

De acordo com o estudo de caso proposto neste trabalho, analisaram-se a usabilidade, missão e os objetivos alcançados por meio de uma aplicação das técnicas de *Data Webhousing* a um s*ite* de Grupo de Pesquisa. Com a aplicação dessas técnicas e com a implementação do *Data Mart* de seqüências de clique, foram feitas as análises em cada caso definido nos requisitos e explorados no *Data Mart*.

6 CONCLUSÕES E RECOMENDAÇÕES

A construção, a disseminação e o acesso às informações em C&T, no País, permitem vislumbrar um conjunto de aplicações voltadas à construção de conhecimento. Em particular, a aplicação de *Data Webhousing* pode contribuir para elucidar relações entre os diversos Grupos de Pesquisa, no Brasil.

Para tal, foram utilizadas técnicas e conceitos da área de *Data Webhousing e Data Warehouse*. Nesse cenário, os dados foram extraídos dos arquivos de *log* de Grupo de Pesquisa e Desenvolvimento da UFSC. Neste capítulo, apresentam-se as conclusões deste estudo que possibilitaram a elaboração deste trabalho, bem como recomendações e as idéias para trabalhos futuros.

6.1 Conclusão

Esta pesquisa contribuiu para desenvolver um estudo de caso prático, que descreveu a maneira de se realizar uma concepção de projeto e como se deve implementar um *Data Mart* de seqüências de clique. Além disso, apresenta uma contribuição para que os Grupos de Pesquisa possam melhorar seus relacionamentos por meio dos seus *site*s *web*. Foi possível fazer algumas reflexões metodológicas do ponto de vista da forma de abordagem do problema, pois a pesquisa pode ser enquadrada como "predominantemente quantitativa", uma vez que requer o uso de recursos e de técnicas estatísticas, que merecem ser destacadas como forma de contribuir para as discussões na área de sistemas de informação. Além do material organizado sobre o trabalho proposto, condensado na revisão bibliográfica, a pesquisa apresenta ainda pontos para reflexão e análise futura, como a questão de identificar o usuário no anonimato e o aprofundamento do pesquisador no modelo proposto.

O mercado de tecnologias de suporte à decisão tem crescido consideravelmente, pois as empresas necessitam cada vez mais de informações estratégicas para criar diferencial competitivo. A geração dessas informações envolve, basicamente, a descoberta de novos padrões nos BDs e a identificação de alternativas para o processo decisório.

A utilização do *Data Webhouse* para sistemas de suporte à decisão está se tornando uma solução cada vez mais comum nas organizações, pelo nível de informações

oferecidas, pela simplicidade de uso e pela facilidade de mesclar ferramentas para sua implementação, tornando a empresa independente de um fornecedor específico.

Pelo fato de ser mais fácil a demonstração de vantagens na utilização de um *DW*, nas áreas de vendas, produtos, seguros e bancos, etc., essas áreas têm sido mais exploradas pelo mercado, apresentando maior número de soluções e implementações. Essa concentração de esforços pode ser vista no grande número de casos de sucesso expostos em eventos especializados no assunto, os quais têm destacado a viabilidade, os benefícios e o retorno de investimentos em *DW*.

Nesse sentido, o *Data Webhousing* apresentado nesta pesquisa pode ser definida como sendo a descoberta de perfis e padrões de acesso aos usuários dos servidores que disponibilizam informação na rede. Como os Grupos de P&D constroem seus *site*s da forma que seus pesquisadores consideram mais apropriada a seus visitantes, a coleta e posteriormente, a análise dos dados referentes aos seus acessos puderam esclarecer a natureza do *site*, auxiliando na compreensão do comportamento dos usuários, de forma a verificar se o *site* está eficientemente projetado e organizado.

Diante do referido estudo, deve-se ressaltar que o processo de pesquisa é um processo puramente humano, por mais informatizado que seja. As informações colocadas diante dos tomadores de decisão são indicadores para melhorar a visualização do contexto que o rodeia. A vantagem de implantação de um *DM* em uma organização é propiciar aos tomadores de decisão uma significativa economia de tempo e esforço no processo de decisão. As decisões continuarão sendo tomadas, como de praxe; a mudança, na realidade, ocorre no grau de certeza que antecede a tomada de decisão e, por conseguinte, sua probabilidade de acerto e precisão podem possuir mais qualidade.

6.2 Objetivos

O primeiro objetivo, ou seja, a exposição da teoria *Data Webhouse*, foi atingido conforme se pode observar no capítulo 3 deste trabalho. Contudo, este capítulo não se limitou apenas na apresentação da teoria, pois procurou-se também enfatizar as técnicas de como gerenciar um *Data Webhouse,* e assim foi possível entender como as organizações coletam dados valiosos sobre seus clientes, os quais podem auxiliá-la na criação de melhores serviços. Além disso, forneceu-se uma visão mais abrangente sobre *Data Webhousing,* tratando de aspectos; por exemplo, de que forma e como se processa o

monitoramento de *site*s por meio de seqüência de clique e quais os principais conceitos atuais que envolvem a área, tanto nos sistemas de informações, quanto nos sistemas gerenciais.

No capítulo 3, seção 3.9, foram expostas algumas metodologias com a finalidade de auxiliar a implantação de um *Data Mart* de seqüências de clique. Essa revisão que está no terceiro capítulo, visou mostrar, passo a passo, como o usuário identifica quais são os dados possíveis de serem capturados, transformados e analisados, segundo as metodologias apresentadas. O estudo permitiu ainda a constatação de um outro ponto que torna a coleta dessas informações provenientes do *clickstream* interessante, vale dizer a possibilidade de emprego dessa tecnologia para obtenção de informações sobre o comportamento dos usuários que navegam em *web sites.*

O terceiro objetivo encontra-se no quarto capítulo, a partir da seção 4.4, na qual foi implementado um *Data Mart* de seqüência de clique, utilizando como estudo de caso um *site* de Grupo de Pesquisa e Desenvolvimento. Dessa forma, seguiram-se os padrões metodológicos desenvolvidos nos estudos e apresentados no terceiro capítulo, a partir da secção 3.9. Procurou-se evidenciar também as técnicas de implementação, bem como as ferramentas utilizadas para implantar o modelo dimensional proposto. Cabe salientar ainda que a pesquisa foi realizada com o objetivo principal de construir um protótipo baseado num modelo dimensional onde foi possível percorrer, na prática, todas as fases para a criação de um *Data Mart* C*lickstream* e também verificar o potencial das ferramentas utilizadas no processo de implementação do *DM.*

Finalmente, o quarto objetivo, ou seja, proposta de melhorias, descrito no capítulo 5, seção 5.5, em que concretizou-se essa proposta, apresentando-se os indicativos estatísticos da aplicação descrita no capítulo 4. Contudo, de acordo com as informações apresentadas, foram mostradas algumas estatísticas do protótipo que teve como finalidade apresentar possibilidades de como melhorar a estrutura dos *site*s dos Grupos de P&D no Brasil. As estatísticas que foram mostradas são específicas de Grupo de Pesquisa, as quais objetivaram mostrar os acessos que seu *site* teve em determinado período, aos quais vale para todos os Grupos de P&D. Dessa Forma, a pesquisa foi concluída apresentando como sugestão um *site* idealizado para os Grupos de Pesquisa de todas as áreas do conhecimento, permitindo, assim, uma melhor interação com todos os Grupos de P&D, no Brasil.

6.3 Contribuição da Pesquisa para o Conhecimento

Este trabalho possibilitou a aplicação de técnicas *Data Webhousing* em um *site* de Grupo de Pesquisa e Desenvolvimento, com os indicativos estatísticos dessa aplicação, permitiu elaborar uma proposta para melhorar as estruturas dos *web sites* dos Grupos de Pesquisas e Desenvolvimento.

Com relação às técnicas utilizadas (*Data Webhousing e Data Warehusing*), estas demonstram serem úteis na resolução de diversos problemas, e quanto à utilização destas técnicas na aplicação proposta, as mesmas ampliam o conjunto de tecnologia e ferramentas que podem prover subsídios nas estruturas dos *sites* de acordo com as informações analisadas. Levando-se em consideração a maneira de exploração desses dados, estas ferramentas podem validar conhecimentos explícitos, ou produzir novos padrões, auxiliando na tomada de decisão.

Dessa forma, a personalização de *web sites,* através de *Data Webhousing* é uma estratégia que permite o aproveitamento das informações deixadas pelo usuário, com o objetivo de tornar o *site* mais próximo das necessidades do seu público.

Então, diante da visão apresentada, o objetivo desta pesquisa foi trabalhar as técnicas apresentadas acima, como meio de melhorar o desenvolvimento da pesquisa de C&T no Brasil. Conseqüentemente, com o resultado estimado deste trabalho contribuiu-se para que os Grupos de Pesquisa e Desenvolvimento conheçam melhor os usuários de seu *site* institucional.

6.4 Limitações do Trabalho

Embora esta pesquisa se apresente como uma proposta viável e mesmo como uma ferramenta útil para melhorias de *site*s de Grupos de P&D, devem-se destacar alguns pontos limitantes na realização e nos resultados do trabalho, tais como:

- dificuldade nas capturas de *logs*, pois o mesmo encontra-se configurado no servidor da organização;
- os resultados alcançados, ainda que satisfatórios são preliminares, tornando-se necessários novos estudos. Estudos estes que devem ser concentrados, tanto em

questões de normalização dos dados, métricas de similaridade e avaliação dos dados, quanto na determinação de novas técnicas de avaliação de dados não estruturados.

- a ferramenta utilizada para a análise das informações foi o Excel 9.0, isto é, uma ferramenta proprietária de determinada empresa que foi utilizada por ser mais fácil a compreensão da mesma; assim, recomenda-se a utilização de outras ferramentas mais "robustas", como Oracle ou IBM. Dessa forma, a implementação e a utilização de outras ferramentas apresentarão resultados mais atuais.

6.5 Recomendações

Sugere-se como recomendações para trabalhos futuros a continuidade deste tema, porém necessita de algumas frente de trabalhos como: ênfase especial na implementação de novas técnicas de estatísticas e de Inteligência Artificial (IA). O aprofundamento nesses aspectos permitirá a elaboração de uma outra ferramenta composta por um conjunto maior de técnicas, que integrem *Data Webhouse* e Raciocínio Baseado em Casos (RBC), possibilitando assim diversas visões e maneiras de coleta de dados, aumentando assim conseqüentemente o suporte às decisões.

Essas pretensões, surgidas de integração *Data Webhousing* e RBC, tem como objetivo a idéia de uma única base de dados que propõe a generalização do mecanismo de busca proposto nas técnicas (RBC para associação de buscas com buscas anteriores), através do acréscimo de outras camadas, de forma a possibilitar o uso em bases de dados com conteúdo e padrões diversos e a apresentar o resultado de forma configurável e integrável aos *sites* dos usuários.

Uma outra sugestão para trabalhos futuros, é a continuidade do modelo dimensional proposto nesta dissertação que leva em consideração o seguinte aspecto; deve-se aperfeiçoar mais o modelo dimensional realizando o mesmo estudo em outras IES, focalizando-se o perfil de cada usuário como um *Data Mart Clickstream,* identificando o caminho percorrido dentro do *site* até a página final em que o usuário finalmente abandona o *site*.

Propõe-se, então, um estudo pormenorizado, no intuito de se produzir um modelo único, cujo *site* seja personalizado de acordo com o gosto e propensão do usuário. Dessa forma, a criação desse modelo, possivelmente virá a aumentar as possibilidades de análises e

de cruzamento de informações. Finalmente, como se observa a continuidade do trabalho se dá por diversos caminhos, dependendo dos indicativos estatísticos que o gestor necessita.

7 Referências Bibliográficas

ALMEIDA, Rubens Queiroz, "O que são Cookies e como Funcionam**"** Online. Disponível em: http://www.dicas-l.unicamp.br/dicas-l/19970711.shtml. Acessado em Dez. 2001.

ABITEBOUL, Serge. "Querying Semi-Structured Data". In Proceedings of Sixth international Conference on Database Theory, pages 1,18, Delphi, Greece, 1997.

BARBOSA, D. M; Sell D; Freitas Jr. O. G; Pacheco R. C. S. "**Uma Metodologia para Desenvolvimento de Data Webhouse Voltado para os Portais Corporativos**", Anais do congresso KmBrasil. UFSCAR – SP. 2002.

BARQUINI, Ramon. "**Planning and designing the Warehouse**". New Jersey, Prentice-Hall, 1996. 311p.

BATINI, C., LENZERINI, M. "**Comparative Analysis Of Methodologies For Database Schema Integration**", ACM Computing Surveys. New York, v.18, nº 4, pág.323 - 364, dez/1996.

BERRY, Michel J. A., LINOFF, Gordon. "**Data mining techniques - for marketing, sales, and customer suppor**", t. John Wiley & Sons, New York, 1997.

BIGUS, Joseph P. "**Data mining with neural networks: Solving business problems from application development to decision suppor**", t. Computing McGraw-Hill, New York, NY, 1996.

BRETZKE, Miriam. "**Marketing de Relacionamento e Competição em Tempo Real com CRM**". Editora Atlas. 2000.

BRETZKE, Miriam. "**O Conceito de CRM Viabilizando o Marketing de Relacionamento para Competir em Tempo Rea**l", Online disponível na Internet, http://www.bretzke-marketing.com.br/abert-cases.htm. Setembro 2001a.

BRETZKE, Miriam. "CRM é mais que uma Tecnologia é Principalmente uma Decisão Estratégica", Online disponível na Internet, online disponível em http://www.bretzke-marketing.com.br/abert-cases.htm. Acessado em Setembro 2001b.

BRETZKE, Miriam. "Estratégia de Marketing de Relacionamento que Realmente Trazem Resultados", Online disponível na Internet, http://www.bretzke-marketing.com.br/abert-cases.htm. Acessado em Setembro 2001c.

CABREIRA, Marcelo Garcia. "Data Mart para Data Webhouse". Online. Disponível em: http://www.data*Web*house.com.br/art_DMFC.htm. Acesso em Jun. 2001, Brazilia 2001.

CAMPOS, Marcelo Luiz "**A Gestão Participativa como uma Proposta de Reorganização do Trabalho em Sistema de Produção Industrial: Uma proposta de Ampliação da Eficácia sob a Ótica da Ergonomia**". Florianópolis, 2000. Dissertação (Mestrado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC-2000".

CIELO, Ivã. “Arquiteturas OLAP”, São Paulo, Maio. 2001. Disponível em: <http://www.datawarehouse.inf.br/>. Acesso em: Mai. 2001.

DALFOVO, Oscar. “**Quem tem informação é mais competitivo**”. Blumenau: Acadêmica, 2000a.

DALFOVO, Oscar; FRANCO, Cristiano Roberto. “**Sistemas de Informação Baseado em Data Warehouse Aplicado a Area Ambiental. In: Simpósio Catarinense de Computação”,** Itajaí. Anais...Itajaí: UNIVALI, 2000. p. 469-479, 2000b.

DEAN, Herington. “Implantando CRM com sucesso”, Online disponível na Internete em call Center.inf.br. www.callcenter.inf.br, Acessado em Dezembro 1999.

DIAS, Cláudia A. “**Portal Corporativo: conceitos e características**”. Ciência da Informação, v. 30, n. 1, p. 50-60, jan./abr. 2001.

DOMENICO, Jorge Antonio Di, “**Definiçao de um ambiente de Data Warehouse em uma Instituição de Ensino Superior**”, Dissertação de mestrado, UFSC – Florianopolis 2001.

ELLY, T. J., *Dimensional Data Modeling*. Disponível em: <*http://www.sybase.com*> Acesso em: 10 out. 2003.

FARIAS, Luciana “**Implementação de um Data Webhouse Simples para Funcionar como Repositório de Informações para Clickstream Analysis**”, Monografia, UFBA – Salvador 2002.

FAYYAD, U. M. “**Data mining and knowledge discovery: making sense out of Data**”, IEEE Expert, 1996a.

FAYYAD, U. M, PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic. From “**Data Mining to Knowledge Discovery: An overview. In: Advances in Knowledge Discovery and Data Mining**”, AAAI Press / The MIT Press, MIT, Cambridge, Massachusets, and London, England, p.1-34. 1996b.

FERREIRA, Sueli Mara S. P. “**Sistema on-line de informação e comunicação / Portal USP**” Universidade de São Paulo: relatório final. São Paulo: USP, 2001.

FIGUEIREDO, A. M. C. M. “**Molap x Rolap: Embate de Tecnologias para Data Warehouse, Developers**” Magazine, ano 2, n. 18, p. 24-25, fev. 1998.

FONTES, E. “**Protegendo a Informação: Fator Crítico para o Negócio**”. Developers’ magazine, ano 2, n. 18, p. 32-33, fev. 1998.

FREITAS, Jr. O. G; Barbosa D. M; Todesco J. L; Pacheco R. C. S. “Abordando o Uso da Orientação a Objetos em um Data Warehouse”, anais CBCOMP2002 – Univali SC. 2002a.

FREITAS, Jr. O. G; Barbosa D. M; Barros E.; Pacheco R. C. S. “**Um Modelo de Gestão do Conhecimento para Aplicação nas Instituições de Ensino Superior**”, anais do Congresso KmBrasil – UFSCAR - SP; 2002.

FURLAN, José Davi; IVO, Ivonildo da Motta; AMARAL, Francisco Piedade. "**Sistema de informações executivas**". São Paulo: Makron Books, 1994.

GATES, B. (1995). "**A Estrada do Futuro**". São Paulo, Companhia das Letras. 1995

GARTNER, Group. Knowledge Management Scenario. Online. Documento capturado em 15/9/2000. disponível na Internet via www.gartnergroup.com 2000

GIL, Antonio Carlos. "**Como elaborar projetos de pesquisa".** São Paulo: Atlas, 1991.

GONÇALVES Klausner Vieira; Ângela Cristina de Oliveira "Data Webhouse", Instituto de Computação – UNICAMP Mestrado Profissional em Computação. 2001

GRUPO STELA, PPGEP – UFSC. "Plataforma Lattes" Periódico documentando a Plataforma Lattes do CNPq. Grupo de Pesquisa e Desenvolvimento do PPGEP da UFSC - 2001.

HAIR Jr., Joseph F., ANDERSON, Rolph E., TATHAM, Ronald L., BLACK, William C. "**Multivariate data analysis**", Prentice-Hall, Upper S. River, 5. ed., N.Jersey, 1998.

HARJINDER, G. e RAO, P. C. "**The Officil Guide to Data Warehousing**". Que Corporation, 1996.

HARRISON, Thomas H. "**Intranet data warehouse**", São Paulo, Berkeley Brasil, 1998.

INMON, William H. "**Como Construir o Data Warehouse**". Rio de Janeiro: Campus, 1997.

INMON, W.H.; WELCH, J.D. e GLASSEY, Katherine J. "**Gerenciando Data Warehouse**". São Paulo – SP, Makron Books, 1999.

INMON, William H., "An architecture for managing clickstream tream data". Disponível em: http://www.billinmon.com/library/library_frame.html. Acesso em Dez. 2001.

INTELLIGENT, "Ralph Kimball" Disponível em www.intelligenteenterprise.com/992112/*Web*house.shtml. Volume 2 Number 18 Acessado em 21 de Dezembro de 1999.

KIMBALL, R. "**Dealing with Dirty Data. DBMS Magazine**"; September 1996. http://www.dbmsmag.com/9609d14.html (05 Jan. 1998). 1996.

KIMBALL, Ralph. "**The Data Warehouse Toolkit. John Wiley & Sons Inc**", New York, 1996b.

KIMBALL, R. "**Data Warehouse Toolkit**" São Paulo, Makron Books, 1998.

KIMBALL, Ralph; REEVES Laura; ROSS Margy; THORNTHWAITE Warren. "**The Data Warehouse Lifecycle Toolkit**": Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons Inc., New York, 1998b.

KIMBALL, R.. "**Data *Web*house**". Tradução de Edson Furman kiewicz e Joana Figueiredo. Rio de Janeiro: Campus, 2000.

KIMBALL, Raph "Working in the *Web* Time" Intelligent Enterprise Magazine. Disponível em: http://www.intelligententerprise.com/db_area/archives/1999/991611/warehouse. Acesso em Novembro. 2001.

KIMBALL, Ralph. "Clicking with your Customer". Online. Disponível em: http://www.intelligententerprise.com/990501/warehouse.shtml. Acesso em Janeiro de 2000b.

KIMBALL, Ralph e Joe Caserta " Clickstream Data Mart ". Online. Disponível em: http://www.intelligententerprise.com/011205/418warehouse1_1.shtml. Acesso Dez. 2001c.

KIMBALL, Ralph "The Special Dimensions of the Clickstream". Disponível em: http://www.intelligententerprise.com/000120/*Web*house.shtml. Acesso em Jan. 2000d.

KIMBALL, Raph e Richard Merz. "**The Data *Web*house Toolkit, Building the *Web*-Enabled Data Warehouse**" . John Wiley & Sons, Inc. 2000. ISBN 0471-37680-9. 2000e.

KIMBALL, Ralph e Richard Merz. "**Data *Web*house, Construindo o Data Warehouse para *Web* **", Ed. Campus, 2000f.

KIMBALL, Ralph, "Clicking with Your Customer", Intelligent Enterprise. Disponível em http://www.intelligententerprise.com/990501/warehouse.shtml, acesso em Jan de 2001.

KIMBALL, Ralph, "The Special Dimensions Of the Clickstream", Intelligent Enterprise, Disponível em http://www.intelligententerprise.com/000120/*Web*house.shtml. Acesso janeiro de 2001.

KIMBALL, Ralph, "**The Data Warehouse Lifecycle Toolkit**", Ed. Wiley, 1998.

KONDRATIUK, E. R. "**Data Warehouse: Detalhes que Fazem a Diferença**". Developers' Magazine, ano 2, n. 18, p. 22, fev. 1998.

KRUGER, Frédi "**Distribuição de Processamento: Fator Crítico para a Escalabilidade de Servidores W*EB***", Monografia, Universidade do Vale Rio dos Sinos – Porto Alegre 2001.

LAMBERT, B. "Break Old Habits To Define Data Warehousing Requirements". Data Management Review; disponível em http://www.data-warehouse.com/resource/articles/lamber11.htm (26 Dez. 1997). Acesso em Dezembro 2001.

LEITÃO, C. Nolla. **"Construção de Aplicação com o Uso de Ferramentas Olap**" Universidade Federal do Rio de Janeiro, Monografia. RJ, 2000.

LIEB, Eric. "A Nova Revolução que se Avisinha", Online Call Center.inf.br. www.callcenter.inf.br, Acesso em junho de 2002.

LITTLE, J. D. C. "**Models and Managers: The Concept of a Decision Calculus**". Management Science, vol. 16, n. 8, p. B466-485, April.1975.

MACHADO, Felipe Nery Rodrigues "**Projeto de Data Warehouse - Uma visão Multidimensional**", Editora Érica - 2000.

MANNI, L. C.; DORSA, L. F. A. "Data Warehouse: Gerenciando a Qualidade dos

MANZONI Jr., R. "**O segredo da produtividade está no uso da informação**". Computerword, 28 a 30 Abr., p. 10-11. 1997.

McKENNA, R. "**Relationship Marketing: Successful Strategies for the age of the customer**", 1ª ed., N.Y, Addison - Wesley Publishing Company, Inc., November, 1991.

McClain, Duncan S., "**Customer Data Integration: The Essencial Component of Effective CRM**", DM Review, June 2000.

MORTON, M. S. S. "**Management Decision Systems: Computer-Based Support for Decision Making**". Boston, Division of Research, Graduate School of Business Administration, Harvard University. 1971.

MENA, Jesus, "**Data Mining your *Web Site***", Copyright © 1999 by Jesus Mena. ISBN # 1-55558-222-2. Excerpted by permission of Digital Press. All rights reserved. Digital Press, 1999.

McKie, Stewart, "**CRM: Customer Role Management**", Intelligent Enterprise, March 2000.

NOGUEIRA, Marcos Diego. I-biznet - Internet Business Network, "Entendendo os Dados de Log (log data)" Online. Disponível em: http://www.i-biznet.com.br/logg/logg20010102182650.asp. Acesso em Dez. 2001.

NIELSEN, Jacob. "**Projetando *Web site*s**". Rio de Janeiro: Campus, 2000.

NIMER, F.; SPANDRI, L. C. "**Obtendo Vantagem Competitiva com o Uso de Data Mining**". Developers' Magazine, ano 2, n. 18, p. 30-31, fev. 1998.

OLIVEIRA, Adelise G. de. "**Data Warehouse Conceitos e Soluções**". Florianópolis: SFO Gráfica e Editora Ltda, 1998.

OLIVEIRA, Roge. "Tendência de Metadados para Data *Web*housing" artigo. Online. Disponível em: http://genesis.nce.ufrj.br/dataware. Acesso em Set. 2001. UFRJ – Rio de Janeiro 2000.

PACHECO, R. C. S., BARCIA, R. M. Plataforma Lattes: **Desenvolvimento de sistemas de informações gerenciais, integrados ao novo modelo de gestão do *CNPq***, 1999.

PEREIRA, M. J. L. B.; FONSECA, J. G. M. "**Faces da Decisão: As Mudanças de Paradigmas e o Poder da Decisão**", São Paulo, Makron Books. 1997.

PINHO, Paulo "Setor de Marketing da Empresa NCR do Brasil", capturado no *site*: www.ideti.com/DataWarehouse/Expositoressp.htm, acessado em março de 2002.

POWER, D. "A Brief History of Decision Support Systems". Disponível em http://power.cba.uni.edu/isworld/dsshistory.html. Acessado em 2002.

PRATES, Maurício. “Os Sistemas de Informações e as Modernas Tendências da Tecnologia e dos Negócios”, Campinas, Abr. 1999. Disponível em: <http://www.puccamp.br/~prates/sistend.html >. Acesso em: Abr. 2001.

TAURION, C. “**Data Warehouse: Estado de Arte e Estado de Prática**”. Developers’ Magazine, ano 1, n. 6, p. 10-11, fev. 1997.

TIEZZI, G. “**O Planejamento Estratégico da Informação**”. Developers’Magazine, ano 1, n. 6, p. 24-25, fev. 1997.

RAPP, Stan & COLLINS, Thomas l. “The new maximarketing”, 1st. Ed. McGraw-Hill, RIBEIRO, Alessandro Coelho e Maria Luiza Campos. “ Gerenciando a Infra-Estrutura de Serviços *Web* Através da Tecnologia de Data Warehousing “ Artigo, Rio de Janeiro 2001New York, pg. 252, 1996.

RICHARD, Li e Jon Salz “Clickstream Data Warehousing”. Online. Disponível em: http://arsdigita.com/asj/clickstream/. Acesso em Ago. 2001.

RODRIGUES, Leonel Cezar. “Impactos dos Sistemas de Informações ”, Jornal de Santa Catarina, Blumenau, 30 Jun.1996. Caderno de Economia, p. 2.

ROMÃO, Wesley. “**Descoberta de Conhecimento Interessante em Banco de Dados sobre Ciências e Tecnologia**”. Florianópolis, 2002. Tese (Doutorado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC-2002

SANT0S, Jose Jadmir Gonçalves, **“O Uso de Tecnologia de Descoberta de Conhecimento em Log de Servidores *Web***” Artigo, UCPEL – Pelotas 2001.

SELL, Denilson. “**Uma Arquitetura para Distribuição de Componentes Tecnológicos de Sistemas de Informações Baseados em Data Warehouse**”. Florianópolis, 2001. Dissertação (Mestrado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC-2001.

SCHONBERG, M., Cofino, T., Hoch, R., Podlaseck, M., Spraragen, S., "**Measuring Success: E-business intelligence is a complex, yet vital, element to building a strong customer base**", Communications of the ACM USA 2001.

SILBERSCHATZ, Abraham; KORTH, Henry; SUDARSHAN, S. **Sistema de Banco de Dado**s. São Paulo: Makron Books, 1999.

SILVA, E. L.; Menezes, E. M. “**Metodologia de Pesquisa e Elaboração de Dissertação**”. 2 a Edição Revisada. Laboratório de Ensino a Distância, UFSC: Florianópolis, 2001.

SILVA, Sandra Regina. “Tecnologia CRM São Todos Iguais”, Online. Disponível na Internet http://www.teletime.com.br/tecnologia. Acesso em Setembro de 2001.

SPRAGUE, Ralph H. e Hugh J.Watson, **Sistemas de Apoio à Decisão**, Campus. 1991

STONE, Merlin, Neil Woodcock, Liz Machtynger. “**CRM: Marketing de Relacionamento com os Clientes**”, São Paulo: Ed. Futura 2001.

STODDER, Davi. “Marketing Automation at Cisco, Intelligent Enterprise”, (Editorial Suppement), November 2000.

STAIR, Ralph M. “**Princípios de Sistemas de Informações**”. Tradução de Maria Lúcia Lecker Vieira e Dalton Conde de Alencar; revisão técnica de Paulo Machado Cavalheiro e Cristina Bacellar. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A, 1998.

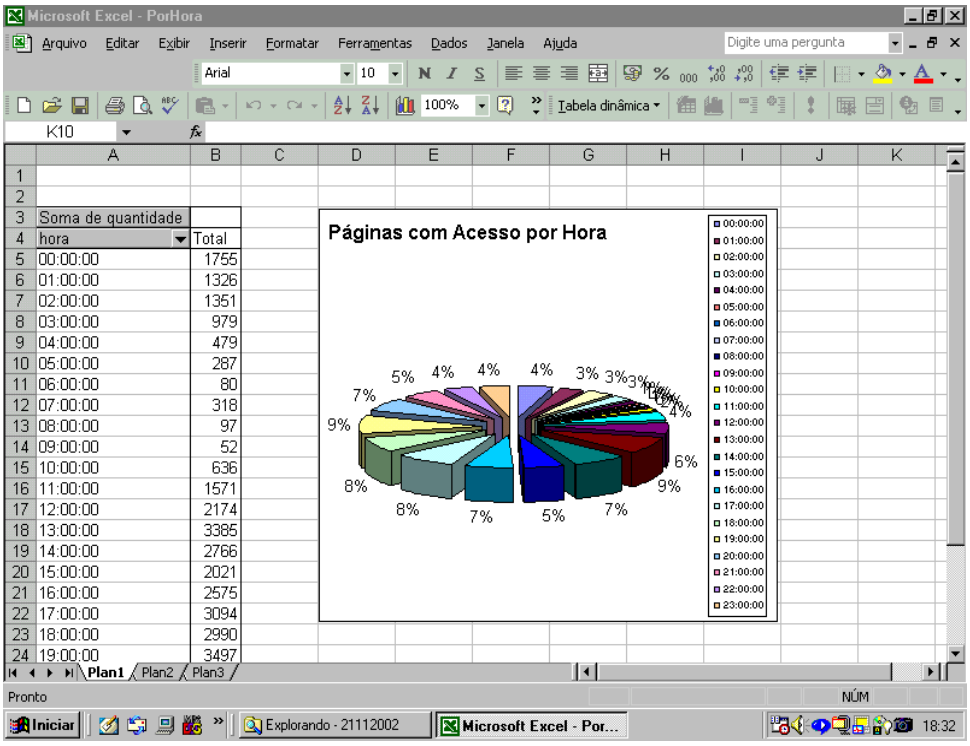
SWEIGER M.; Madsen Mark R.; Langston Jimmy; Lombard Howard “**Clickstream Data Warehousing**” Published by John Wiley & Sons, inc. USA. 2002.

VALENTE, Daphnis Lopes. “**Estudo sobre Armazém de Dados**” CPGCC da UFRGS, Porto Alegre, 1996.

VOELCKER, Ricardo. **Data Webhouse - Clickstream.** Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2001.

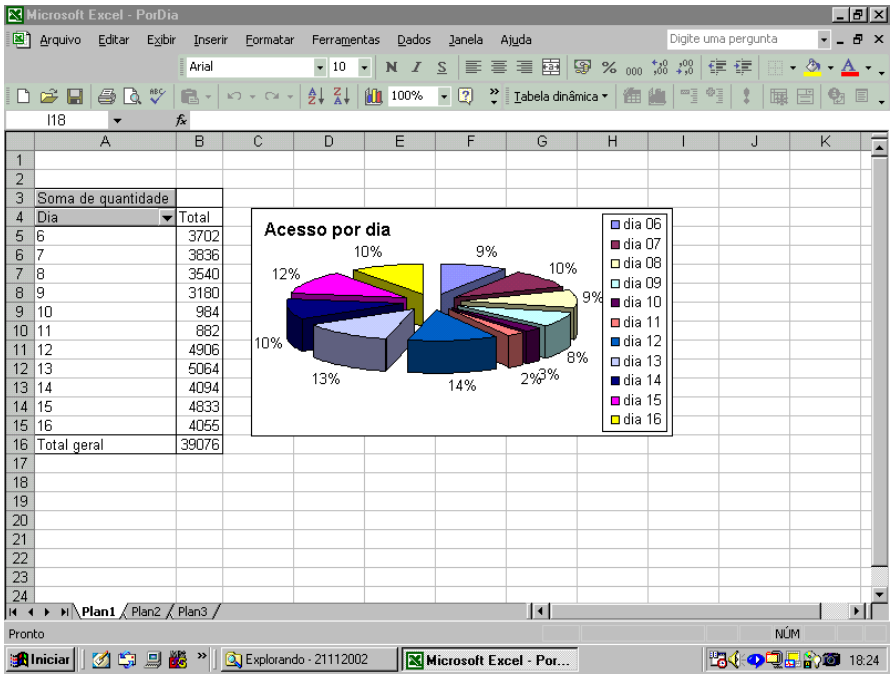
WILEY The publisher, John Wiley & Sons Editorial Review : “**The Data*Web*house Toolkit: Building the *Web*-Enabled Data Warehouse**” 2000.

ANEXO I



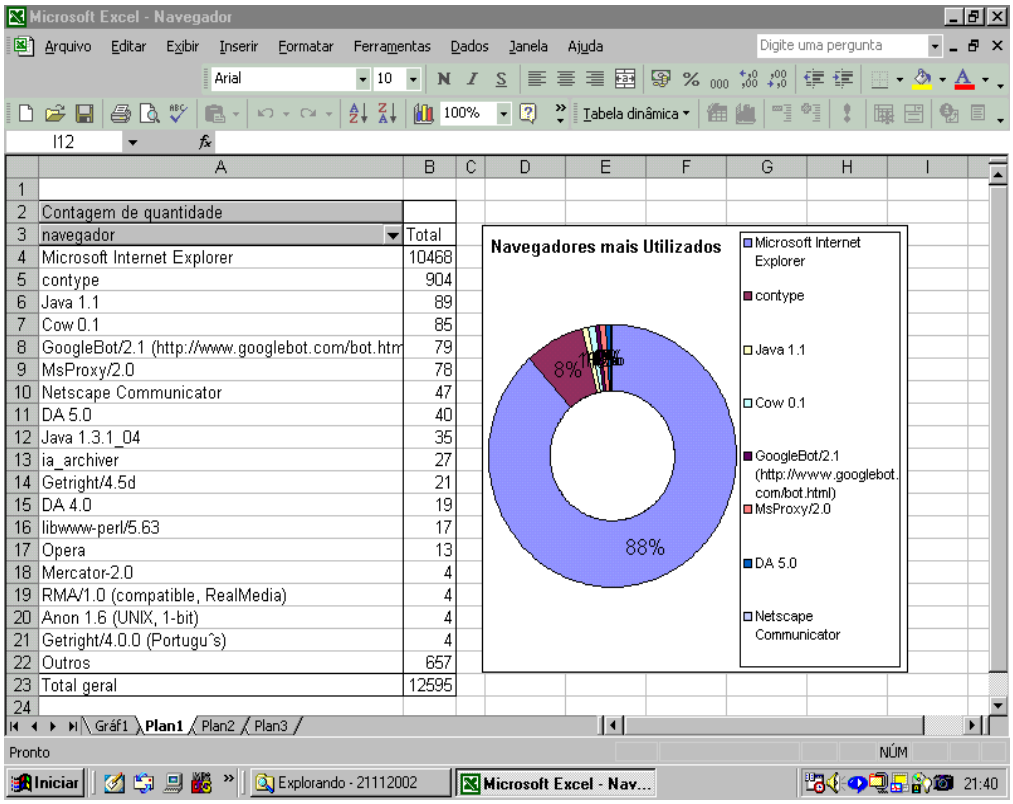
Soma de quantidade	
hora	Total
00:00:00	1755
01:00:00	1326
02:00:00	1351
03:00:00	979
04:00:00	479
05:00:00	287
06:00:00	80
07:00:00	318
08:00:00	97
09:00:00	52
10:00:00	636
11:00:00	1571
12:00:00	2174
13:00:00	3385
14:00:00	2766
15:00:00	2021
16:00:00	2575
17:00:00	3094
18:00:00	2990
19:00:00	3497

GRÁFICO 1 - Qantidade de páginas mais acessadas por hora



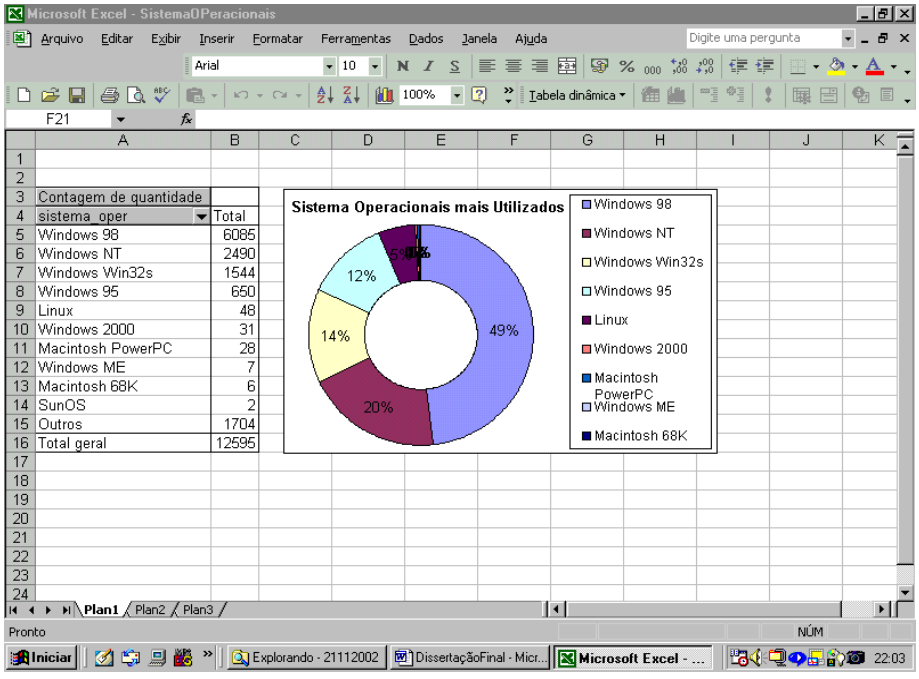
Soma de quantidade	
Dia	Total
6	3702
7	3836
8	3540
9	3180
10	984
11	882
12	4906
13	5064
14	4094
15	4833
16	4055
Total geral	39076

GRÁFICO 2 – Quantidade de páginas acessadas por dia



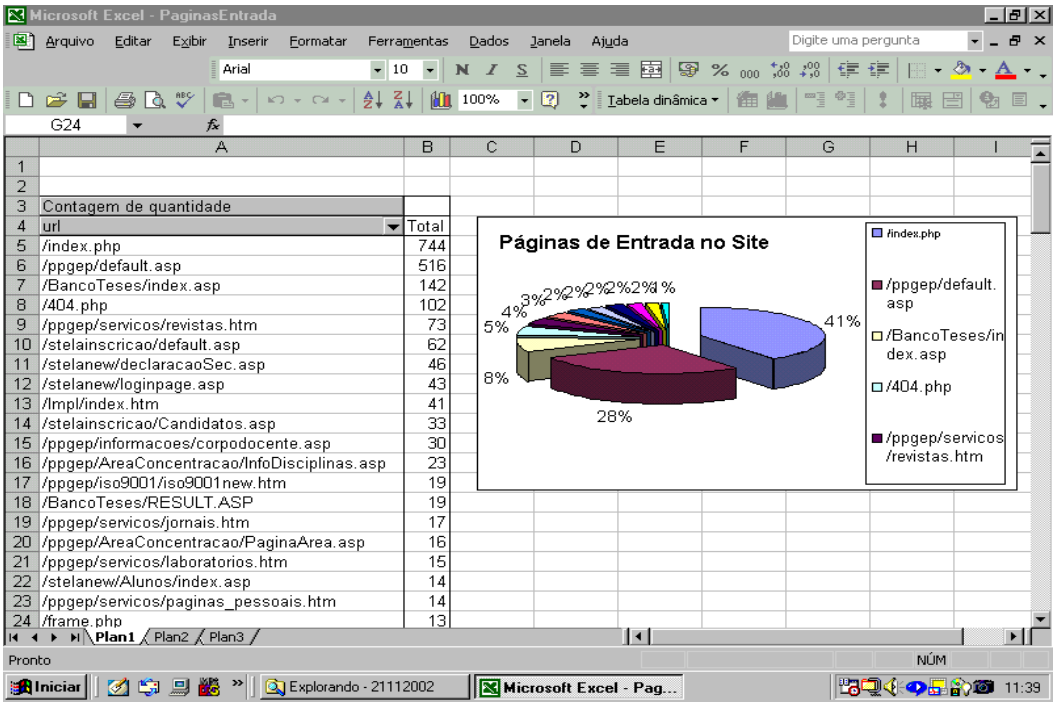
Contagem de quantidade	
navegador	Total
Microsoft Internet Explorer	10468
contype	904
Java 1.1	89
Cow 0.1	85
GoogleBot/2.1 (http://www.googlebot.com/bot.htm	79
MsProxy/2.0	78
Netscape Communicator	47
DA 5.0	40
Java 1.3.1_04	35
ia_archiver	27
Getright/4.5d	21
DA 4.0	19
libwww-perl/5.63	17
Opera	13
Mercator-2.0	4
RMA/1.0 (compatible, RealMedia)	4
Anon 1.6 (UNIX, 1-bit)	4
Getright/4.0.0 (Portugu^s)	4
Outros	657
Total geral	12595

GRÁFICO 3 - navegadores da Internet mais utilizados pelos visitantes



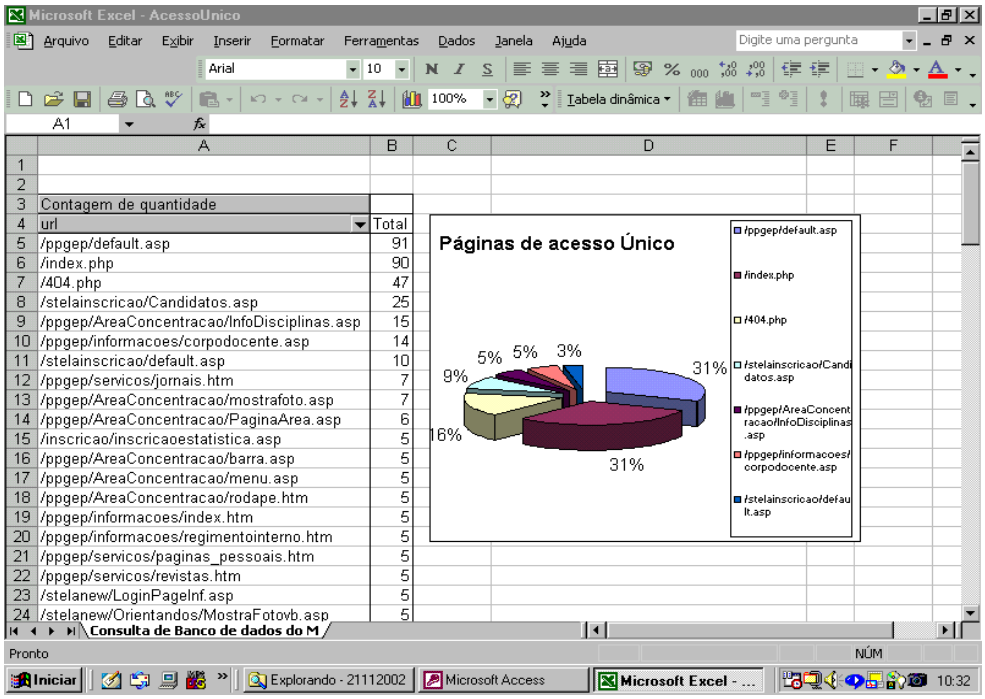
Contagem de quantidade	
sistema_oper	Total
Windows 98	6085
Windows NT	2490
Windows Win32s	1544
Windows 95	650
Linux	48
Windows 2000	31
Macintosh PowerPC	28
Windows ME	7
Macintosh 68K	6
SunOS	2
Outros	1704
Total geral	12595

GRÁFICO 4 - Sistema operacional mais utilizado no site



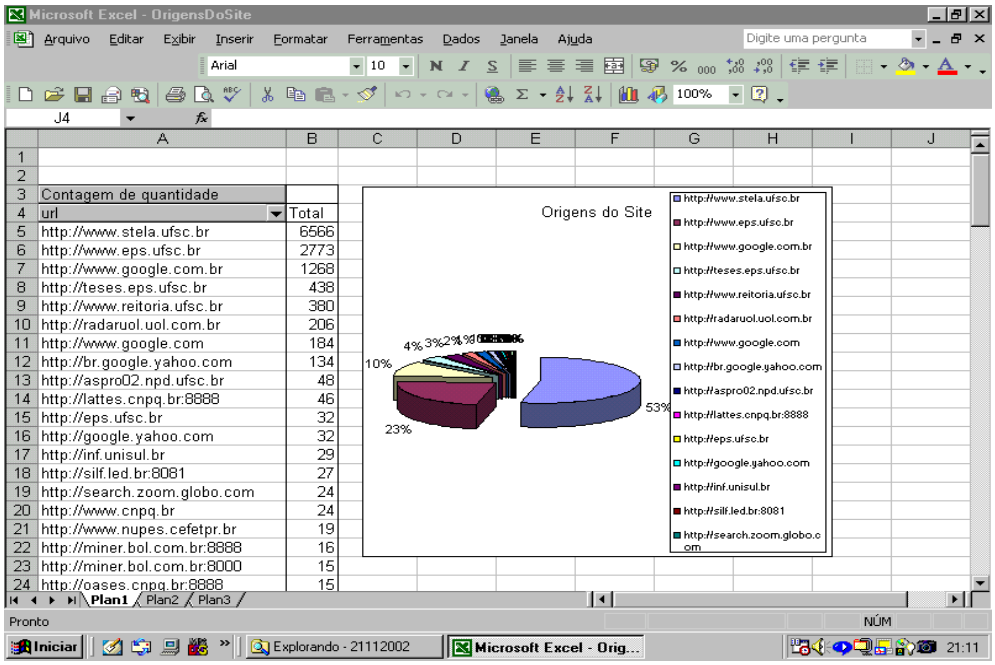
Contagem de quantidade	
url	Total
/index.php	744
/ppgep/default.asp	516
/BancoTeses/index.asp	142
/404.php	102
/ppgep/servicos/revistas.htm	73
/stelainscricao/default.asp	62
/stelanew/declaracaoSec.asp	46
/stelanew/loginpage.asp	43
/Impl/index.htm	41
/stelainscricao/Candidatos.asp	33
/ppgep/informacoes/corpodocente.asp	30
/ppgep/AreaConcentracao/InfoDisciplinas.asp	23
/ppgep/iso9001/iso9001new.htm	19
/BancoTeses/RESULT.ASP	19
/ppgep/servicos/jornais.htm	17
/ppgep/AreaConcentracao/PaginaArea.asp	16
/ppgep/servicos/laboratorios.htm	15
/stelanew/Alunos/index.asp	14
/ppgep/servicos/paginas_pessoais.htm	14
/frame.php	13

GRÁFICO 5 - Páginas de entrada no site



Contagem de quantidade	
url	Total
/ppgep/default.asp	91
/index.php	90
/404.php	47
/stelainscricao/Candidatos.asp	25
/ppgep/AreaConcentracao/InfoDisciplinas.asp	15
/ppgep/informacoes/corpodocente.asp	14
/stelainscricao/default.asp	10
/ppgep/servicos/jornais.htm	7
/ppgep/AreaConcentracao/mostrafoto.asp	7
/ppgep/AreaConcentracao/PaginaArea.asp	6
/inscricao/inscricaoestatistica.asp	5
/ppgep/AreaConcentracao/barra.asp	5
/ppgep/AreaConcentracao/menu.asp	5
/ppgep/AreaConcentracao/rodape.htm	5
/ppgep/informacoes/index.htm	5
/ppgep/informacoes/regimentointerno.htm	5
/ppgep/servicos/paginas_pessoais.htm	5
/ppgep/servicos/revistas.htm	5
/stelanew/LoginPageInf.asp	5
/stelanew/Orientandos/MostraFotovb.asp	5

GÁFICO 6 – Páginas com acesso único



Contagem de quantidade	
url	Total
http://www.stela.ufsc.br	6566
http://www.eps.ufsc.br	2773
http://www.google.com.br	1268
http://teses.eps.ufsc.br	438
http://www.reitoria.ufsc.br	380
http://radaruol.uol.com.br	206
http://www.google.com	184
http://br.google.yahoo.com	134
http://aspro02.npd.ufsc.br	48
http://lattes.cnpq.br:8888	46
http://eps.ufsc.br	32
http://google.yahoo.com	32
http://inf.unisul.br	29
http://silf.led.br:8081	27
http://search.zoom.globo.com	24
http://www.cnpq.br	24
http://www.nupes.cefetpr.br	19
http://miner.bol.com.br:8888	16
http://miner.bol.com.br:8000	15
http://oases.cnpq.br:8888	15

GÁFICO 7 - Origens do site